

Good eggs

There is no place for ageism in reproductive medicine.

Earlier this year, Harvard researchers turned dogma on its head by showing that, in mice, eggs are produced well into adult life (*Nature* **428**, 145–150; 2004). This suggests that the menopause occurs not when eggs produced by the embryo eventually run out, as previously thought, but when cell death outpaces egg production in the adult ovary. If this is true for humans, the implications for treating infertility and health problems related to the menopause are profound. But it would be unfortunate if the suitability of older mothers also became a prominent part of discussions.

It is appropriate to scrutinize the societal impacts of any research, but a double standard may exist with regard to fertility. Men who father children into their golden years are more likely to receive back slaps than pressure to stop. In contrast, many fertility clinics deny services to women over 50, and even as young as 45. The American Society for Reproductive Medicine's ethics committee discourages oocyte donation to postmenopausal women, citing a variety of concerns relating to the medical condition of mother and baby, as well as the psychological effect on the baby of having an older mother.

In truth, the likelihood that any individual will be a good parent depends on a variety of factors, only one of which is age. And while it is true that children of older parents have a higher risk of birth defects, this can be lessened by screening high-risk embryos for genetic defects before implantation. As for the medical risks, a 2002 study showed that pregnancy for women in their fifties from *in vitro* fertilization (IVF) is just as safe as for younger women who had IVF from donor eggs, and

found no reason to exclude women over 50 from pregnancy on the basis of age alone (*J. Am. Med. Assoc.* **288**, 2320–2323; 2002).

Delaying the menopause is usually viewed in the context of fertility research, the subject of this week's Outlook (see page 37). Current research may help to prevent the debilitating conditions that arise when ovaries come to the end of their natural working lives. Even now, women spend nearly half their lives in menopause, an experience that can range from mild discomfort to debilitation. Osteoporosis is associated with loss of ovarian function, and there is evidence that cardiovascular disease, mood disorders, neurological problems and diminished sex drive are also connected. Genetic mouse models in which the menopause is delayed may eventually help disentangle which health risks are a direct consequence of menopause, and which are brought about by ageing (see *Nature Genetics* **21**, 200–203; 1999).

Whether the menopause is seen as a medical condition requiring treatment or a phase of life to be tolerated, it is clear that, as human lifespan is extended, the ability to delay the menopause will be attractive to many women. Hormone-replacement therapy is no alternative to functioning ovaries, so it is to be hoped that a better understanding of ovarian regeneration will lead to improved treatments.

Research into women's reproductive health is important for its potential benefits to quality of life and overall health. It should not be sidetracked by detractors who make the patronizing and outdated argument that older women shouldn't have babies that they will be too old to care for. ■

Bad funding

The research of an environmental regulator is unlikely to win public trust if it relies on money from industrial lobby groups.

The US Environmental Protection Agency (EPA) hasn't violated any laws by accepting US\$2 million from a chemical-industry lobby group to help it conduct toxicology research. But it is crossing a line that a regulator in the world's wealthiest nation should not have to.

As reported on page 6, environmental groups claim that accepting the funds represents a clear conflict of interest for the EPA, which is tasked with regulating companies that are members of the American Chemistry Council, the lobby group in question.

The study under scrutiny will examine the pathways by which young children are exposed to pesticides and household chemicals. The decision to accept the money is driven, EPA officials say, by a desire to conduct as comprehensive a study as possible, given constraints on the agency's research budget, which has remained essentially constant for almost a decade and is expected to dip slightly this year. The lobby group's cash will allow EPA scientists to analyse samples for the presence of potentially toxic compounds that they would otherwise have to leave out of the study. The agreement between the EPA and the chemistry council attaches no constraints, they add.

This is not a unique case: the EPA has struck agreements with dozens of companies to study a vast range of environmental problems over the past decade. But \$2 million is one of the largest sums ever accepted by the agency from an outside source and, in this case,

the source has a direct interest in the outcome. Even if the wording of the agreement protects the research from any direct interference, it is hard for officials to argue that it doesn't give the appearance of compromising the agency's independence as a regulator.

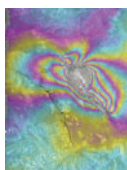
Administrators at the agency may have little choice but to make deals with industry in order to move essential research forward, but there are ways to limit the appearance of impropriety. For example, industry could agree to give to a general research fund, rather than picking and choosing specific projects to support. Or it could partner universities or environmental groups on projects that all parties agree are important. Such moves would allay some of the concerns of environmentalists, and allow industry groups to support the science that they believe will ultimately lead to better regulation.

But above all, it is essential — not only for trust but also for adequate regulation — that the EPA's science budget be increased. The agency's mission requires it constantly to address questions of exposure and toxicity that are politically controversial, scientifically complex and poorly understood. It has been apparent for years that its research budget of about half-a-billion dollars a year is inadequate. Congress must give the cash-strapped agency the resources it needs to properly implement the plethora of mandates that it is meant to enforce. This idea is not so radical — it is supported by industry and environmental groups alike. ■





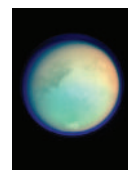
Strain station
Mouse sequencing
project promises
disease insights
p5



Up in the air
US Earth scientists
back call for surface
monitoring by satellite
p6



Treasure hunt
Museum tries to
stop illegal sales of
archaeological finds
p7



Full moon
Cassini captures
atmospheric shot
of Titan
p8

China takes steps to secure pole position in primate research

David Cyranoski, Tokyo

China is poised to become a global centre for biomedical research in primates, specialists from around the world heard at a meeting in Kunming last week.

But some observers are already worried that the circumstances drawing primate researchers to China — low costs, uncertain regulations and the absence of animal-rights groups — could lead to trouble in the future.

Researchers from the United States and Europe are increasingly seeking collaborations with primate centres such as the Kunming Institute of Zoology in southwestern China, which hosted the symposium on 27–30 October.

The symposium promoted biomedical research that uses primates as models. Two dozen talks, mostly given by scientists based in Britain or the United States, covered brain disorders, primate evolution and the use of primates to study cancer, HIV treatments and immune disorders.

“Our research work must be of international standard,” says Weizhi Ji, director of the Kunming Institute. Foreign researchers say that they were awed by the scale of Ji’s facility, which currently hosts 1,400 experimental monkeys, including 300 in isolation. “It’s bigger than anything I’ve ever seen,” says Frances Gotch, an HIV-vaccine researcher who formerly worked with primates in Europe. “This facility undoubtedly could be of great service to the rest of the world.”

Fraser Wilson, a neurophysiologist who until recently was based at the University of Arizona in Tucson, is one of three US researchers who will take a full-time position at Kunming to study neurophysiology. Wilson says that a mountain-top field station there will allow them to track the activity of single neurons in monkeys that are freely exploring large-scale three-dimensional space. “Come here in six months and see what we are doing,” he boasts. “It will be world-class research.”

Other primate-research facilities are also blossoming in China. A collaboration



Unknown quantity: standards for animal care in China’s burgeoning primate laboratories have yet to be established.

between Bruce Lahn, a researcher at the University of Chicago, Illinois, and Sun Yat-sen University in southern Guangzhou, for example, has begun producing transgenic and inbred monkeys that could be used, like their mouse equivalents, to study disease (see *Nature* **424**, 239–240; 2003). And the Institute of Neuroscience in Shanghai will soon start taking functional magnetic resonance images of primates to study their neurophysiology.

Model behaviour

Primates are even making their way into cardiovascular research. Anthony Chan of the Yerkes National Primate Research Center in Atlanta, Georgia, a pioneer of the use of transgenic monkeys, is teaming up with the new Institute of Molecular Medicine in Beijing to create transgenic monkeys that will serve as models for studying atherosclerosis, heart failure and diabetes. Chan, who says that recent studies show the superiority of non-human primates as models compared with other lab animals (G. S. Roth *et al. Science* **305**, 1423–1426; 2004), thinks

that researchers in China are well equipped to pursue such work.

The research will be relatively cheap to do in China, with monkeys costing less than US\$1,000 — about a tenth of what they cost in Europe. And it will be free of the pressure from animal-rights groups, which, according to veterinarian scientist Paul Malatesta of the Robert Wood Johnson Medical School in New Jersey, has prevented his home institution from working with primates. “It’s not worth the expense,” he says, “and you don’t want to be the target of the animal-rights people.” Animal-rights groups have, in the past, criticized Wilson’s experiments at the University of Arizona, and held vigils there in protest.

Ji says that China is fundamentally different from the West in that “human health is always given the first priority and there is less emphasis on animal rights”. But he says that he recognizes the importance of strong ethical regulations — and that the

symposium will help to establish these.

Gotch says that the standards of animal care at Kunming seem to be as strict as those in Britain in terms of emergency exits, cleanliness and operating facilities. “There is nothing ‘third world’ about it,” she says.

Ji claims that China’s regulations on animal research already meet international standards. But in a country where protests or even discussion are rare without government sanction, foreign activists question that. Problems that have arisen with primate care in the United States “will tragically be repeated”, says Elliot Katz, a veterinarian who founded In Defense of Animals, an animal-rights group based in Mill Valley, California, “And they’ll feel there are even fewer limitations.”

Ji concedes that the institution-based review boards that monitor research ethics in the United States have rarely been established in China. But he hopes that a committee on animal care and use — a group that he heads at Kunming — will serve as a model for primate research in the rest of China. ■

Climate change clouds commercial licence to krill

Emma Marris

Marine biologists and government officials meet in Australia this week to agree on safeguards for Antarctic krill.

These pink crustaceans (*Euphausia superba*) darken the Southern Ocean in enormous swarms and underpin vital ecosystems. But according to evidence published in this issue (see page 100), melting ice in the region could compound the threat that the species already faces from commercial fishing.

A research team, led by Angus Atkinson of the British Antarctic Survey, based in Cambridge, suggests that the krill population is declining with the retreat of winter sea ice in the region, which most scientists think is being driven by global climate change. In one region, the krill population has dropped by 80% over the past 30 years.

The Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR), which was set up by the Antarctic Treaty and regulates the Southern Ocean's fisheries, began its annual meeting on 25 October in Hobart, Tasmania. The commission sets annual limits for krill, based on observations of the health and age of the population. But the commission's current monitoring system looks only at predators, not at long-term environmental change. Some marine biologists hope the research findings will change this.

Krill are caught mainly as food for domestic animals and fish, because they are fit for human consumption only when very fresh. The fishery is relatively small at the moment, but most observers expect that to change. Keith Reid, a marine ecologist at the British Antarctic Survey, predicts a "Klondike krill rush" as technology makes harvesting them easier and as other species are fished out. He hopes CCAMLR will be ready. "Here is a stock that has declined dramatically in the absence of fishing, but most models assume you have a stable population."

Denzil Miller, the executive secretary of CCAMLR, hopes to have a new monitoring framework in place within two years — in advance of any boom. Such a framework would take climate change into account, and divide the krill's range into small areas with individual caps to prevent overfishing of single sites.

"If there is a sudden explosion in the fisheries, we must have a framework to manage that," says Miller. "And we have to have these decisions pre-agreed." ■

Early embryos fuel hopes for shortcut to stem-cell creation

Helen Pearson, New York

In an advance that could boost the production of stem cells for medical research, fertility researchers have grown human embryonic stem cells from an embryo that was younger than any used before.

Yury Verlinsky and his colleagues at the Reproductive Genetics Institute in Chicago, Illinois, grew cells from four-day-old human embryos called morulae (see *Reproductive BioMedicine Online* www.rbmonline.com/Article/1558). Previously, stem cells have been grown only from blastocyst-stage embryos, which are a day or two older and have more specialized tissues.

Researchers in the field say the new technique is quicker and simpler than previous methods and could boost the success rate for making embryonic stem-cell lines, which are in short supply. "It takes a huge step out of the whole process," says Peter Braude, who studies human embryonic stem cells at Guy's, King's and St Thomas' School of Medicine in London.

Verlinsky and his team attempted to grow stem cells from 46 human morulae left over from *in vitro* fertilization (IVF) treatment. They succeeded in making eight stem-cell lines — a success rate roughly equivalent to that achieved using blastocysts.

In its paper, the team presents preliminary evidence that these stem cells are able to generate a similar repertoire of cell types to those grown from blastocysts. They have already used both techniques to create a library of stem-cell lines for disease research, including many carrying genetic abnormalities (see *Nature* **429**, 691; 2004).

The new method is advantageous because

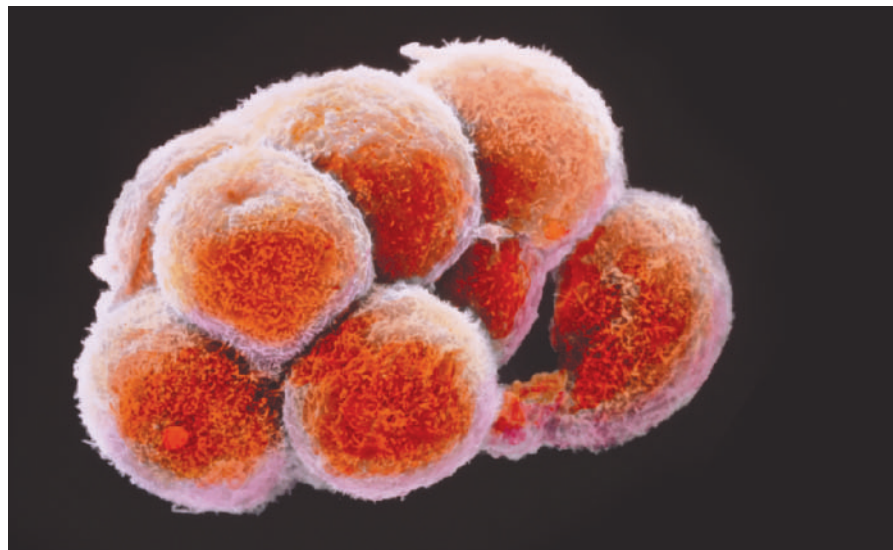
it avoids the delicate task of culturing embryos until they grow into blastocysts, a period when perhaps half of them stop growing or die. It also cuts out a laborious step from the conventional technique, in which the outer shell of the blastocyst is destroyed to gain access to the precious 20–30 stem cells inside. In the new method, the entire morula of around 60–70 cells is grown in culture.

The advance is unlikely to affect the contentious global debate over the ethics of embryonic stem-cell research, scientists say. Opponents of such research point to the fact that stem-cell lines can only be created by destroying a human embryo.

But Braude notes that researchers will now want to test the more radical concept of growing stem cells from a single cell plucked from a morula-stage embryo without damaging the rest of it. That would also raise the prospect that couples having IVF treatment could ask to have one of their embryo's cells removed before it is implanted in the womb. The cell could be grown into a stem-cell line and stored in case the child should need it for disease therapy in the future.

Because this technique would avoid destroying the entire embryo, stem-cell researchers say that it could also serve to quell some of the ethical arguments against the use of human embryos for research.

Researchers might also use the new method to grow stem cells from morula-stage embryos that have stalled in their development and are incapable of growing into babies, suggests stem-cell researcher Jose Cibelli of Michigan State University in East Lansing. "You could remove a big obstacle from the ethical standpoint," he says. ■



More from morulae: stem-cell lines have been created from embryos consisting of very few cells.

P. M. MOTTA, J. VAN BLERKOM/SPL

Mouse sequencing plan aims to boost models

Helen Pearson, New York

Efforts to find genes involved in human disease are set to benefit from a plan to sequence the genetic code of 15 different mouse strains.

Lab mice might look alike, but in fact scientists work with at least 100 different inbred strains. These strains differ hugely in their behaviour, physiology and susceptibility to disease, and researchers are keen to find the differences in the genetic sequences that underlie these traits, with the aim of locating counterparts in humans.

The sequencing project plans to decipher the genomes of 15 mouse strains within two years. It will be largely funded by the US National Institute of Environmental Health Sciences and the sequencing will be done by Perlegen Sciences, based in Mountain View, California. "People are delighted," says mouse geneticist Kenneth Paigen of the Jackson Laboratory in Bar Harbor, Maine.

Researchers have long puzzled over the differences between mouse strains (see *Nature* 415, 8–9; 2002). After a few weeks on a high-fat diet, for example, one strain's cholesterol level hovers at a healthy 100 milligrams per decilitre, whereas another's rockets to more than five times that amount.

Tracking down the genes involved in conditions such as high cholesterol, diabetes, obesity or cancer is a long process. Researchers cross a mouse strain susceptible to diabetes, for example, with a disease-resistant strain, and identify the chunks of the genome that are inherited with the disease. They then laboriously scour the DNA in these regions for 'susceptibility genes'.



Take the strain: a catalogue of genetic differences between mouse breeds will aid disease research.

The new line-up of mouse genomes should accelerate the process significantly. The genomic data will show researchers exactly how the DNA of the diabetes-prone strain differs from that of other strains. So researchers should be able to home in quickly on relevant genes and perhaps find corresponding ones involved in human disease.

Geneticists have had access to the genome of one widely used mouse strain, called C57BL/6J, since 2002 — and, for a fee, genomes of a few other strains are available from Celera Genomics in Rockville, Maryland. The project's organizers say this is the most comprehensive attempt yet to analyse the genomes of different strains, and that the code will be deposited in public databases.

The 15 strains were selected because they are widely used in research, are evolutionarily diverse and show different

propensities to common diseases.

Researchers hope to link their genetic information to the Mouse Phenome Project, an international effort launched in 2000 to collate the anatomical, physiological and behavioural differences between mouse strains. This should help researchers to choose the most appropriate strain for studying a disease.

Ultimately, the amount of detailed biological and genetic data collected might be so great that it eliminates the need for many mouse crosses altogether, says the leader of the Phenome Project, Molly Bogue, also at the Jackson Laboratory. Researchers could simply select a panel of mouse strains with different cholesterol levels, and use computer analyses to help pinpoint the candidate genes that differ between them. "That's what we're hoping for," she says. ■

Green groups baulk at joining nanotechnology talks

Kendall Powell

One of the first efforts to get parties round the table to discuss nanotechnology has got off to a faltering start, after leading environmental groups declined to take part.

The International Council on Nanotechnology (ICON) has been set up to drive open discussion about the benefits and pitfalls of the field, which comprises a clutch of technologies involving materials and components on the scale of a billionth of a metre. The council held its first meeting on 28 October at Rice University in Houston, Texas, which has strong research programmes in the implications of nanotechnology.

But the three main environmental groups invited to participate said they were not ready to do so, complaining that the council was likely to be biased because it depended on industry funds.

Jennifer Sass of the National Resources

Defense Council and Scott Walsh of Environmental Defense, both based in Washington DC, and Pat Mooney of the Canadian ETC Group, based in Ottawa, all turned down invitations to join ICON. Walsh and Mooney participated in last week's meeting as guests.

Researchers and industrial backers of nanotechnology hope that discussion of its effects will help avoid the public mistrust that has plagued fields such as agricultural biotechnology.

"We welcome anyone who thinks they have a stake in this discussion," says Kristen Kulinowski, director for education and public policy at Rice's Center for Biological and Environmental Nanotechnology, which manages ICON. Kulinowski says the council discussed how to make its decision-making truly independent of its funding sources.

Sass would prefer the council to be

publicly funded. Mooney says he would join if ICON included more members from the academic world, from trade unions and from developing countries.

But academic members say the current arrangements are satisfactory. "I'm not concerned about industry sponsorship," says Günter Oberdörster, a toxicologist at the University of Rochester in New York, who studies nanoparticles. "I am concerned about non-governmental organizations possibly not being part of it," he says. "We should find the answer together, in a transparent process."

William Provine, a council member who tracks nanotechnology applications for DuPont, the chemicals company based in Wilmington, Delaware, agrees. "We need to establish a credible group of stakeholders," he says, "and we have not gotten there, yet." ■

EPA accused of conflict of interest over chemicals study

Geoff Brumfiel, Washington

The Environmental Protection Agency (EPA) is under fire from a watchdog group for accepting US\$2 million from a chemical-industry lobby organization to study the effects of pesticides and household chemicals on children.

The Environmental Working Group, a Washington-based non-profit body, says the funding represents a clear conflict of interest for the EPA. "The concern is that the regulated industry, which makes the products being tested, is buying access to the study's design, data and findings," says Jane Houlihan, the group's vice-president for research.

But Paul Gilman, assistant administrator for research and development at the EPA, says the money comes with "no strings attached". He adds that the cash-strapped agency would be unable to do the research without the money. "Is ignorance better?" he asks.

The dispute concerns the Children's Environmental Exposure Research Study, a three-year investigation that was originally designed to look at children's contact with commercial pesticides. When EPA scientists wanted to expand this to include chemicals commonly found around the home, they approached the American Chemistry Council (ACC), an industry lobby group representing 135 chemical manufacturers, according to Carol Henry, the council's vice-president for research.

Henry says that the ACC agreed to give an additional \$2 million to include chemicals found in everything from furniture coatings to cosmetics.

The ACC was willing to fund the study because it believes it will ultimately lead to sounder regulations, says Henry. "If you don't know how chemicals get to various individuals, you have to make a lot of assumptions under the regulatory procedures," she says.

Gilman adds that although the industry group will have a right to review the study 45 days before publication, it will have no influence on the research findings. "We're being very open and public about this process," he says.

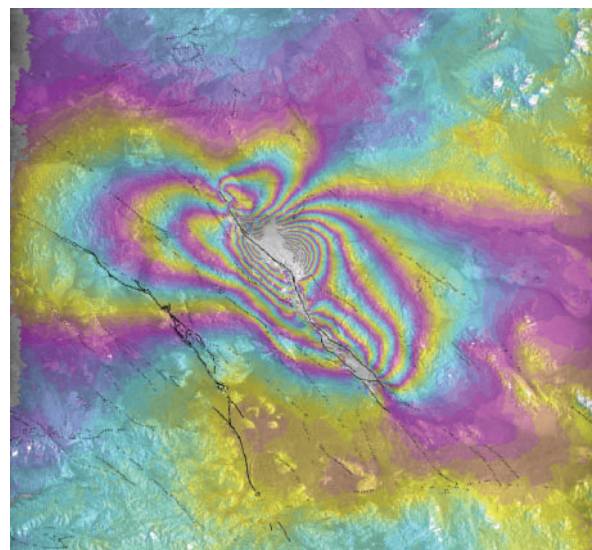
But Houlihan remains concerned. "The EPA's research budget is over \$500 million a year, so you have to ask why the agency is relying on the lobbying arm of its own regulated industries for such a small sum of money in a study that could be critical to children's health protection," she says. ■

Standing room only signals US zeal for Earth imaging

David Cyranoski

Earth-science researchers in the United States are clamouring for a satellite system dedicated to observing the Earth's surface. Such a tool, they say, will transform studies of earthquakes, volcanoes, ice sheets and global land cover.

Organizers of a workshop to discuss the issue in Oxnard, California, last month were forced to turn people away. Some 350 researchers wanted to attend, 300 more than expected. Yet despite this growing scientific support, the price-tag of US\$450 million means that the project has so far failed to get off the ground.



When the Earth moved: radar interferometry tracked ground displacements during California's 1999 Hector Mine earthquake.

The satellite-based technology that the researchers want is called InSAR, for interferometric synthetic aperture radar. These instruments bounce radar off the ground, creating a precise map of the surface, including information about the type of ground cover. Unlike global positioning systems, which can only take spot measurements of the location of receiver stations, InSAR-equipped satellites can image large continuous swaths of land.

"This is critical for studying natural hazards, polar ice sheets and large-scale land-cover changes resulting from climate change," says Jean-Bernard Minster, a geophysicist at Scripps Institution of Oceanography in La Jolla, California.

The few InSAR satellites launched by Japan, Europe and Canada have made some eye-opening discoveries in the past decade. Volcanoes thought to be dormant have been shown to be active, for example, and active volcanoes have been shown to be inflating

and deflating much more frequently than previously thought. Massive ice sheets in Antarctica and Greenland have also been mapped.

But researchers say they cannot continue to rely on these satellites for InSAR data, as many were not optimally designed for radar imaging. Some have imprecise orbits, and so do not pass over the same spot of land on each revolution, making it difficult to measure change. Others use radar wavelengths that are unable to penetrate vegetation to image the ground beneath.

The lack of suitable satellites means that many events important to Earth scientists

were not imaged by InSAR equipment. The area struck by California's 1994 Northridge earthquake, for example, was only imaged two months before and two years after the event, says Andrea Donnellan, a geophysicist at NASA's Jet Propulsion Laboratory in Pasadena, California. And the recent Parkfield earthquake (see *Nature* 431, 618; 2004) was not covered at all, she adds. "Missed opportunities are too many to mention," says Minster.

After ten years of requests and failed grant applications, the US InSAR hopefuls still do not have a dedicated satellite. Donnellan, who convened last month's workshop, hopes this will soon

change. The group is currently applying to the US space agency NASA for funding that could kick-start an InSAR programme. But the lengthy grant-approval process and the design and planning procedure mean that the earliest possible launch would be in 2010.

Donnellan says that a dedicated InSAR system would quickly prove its worth, in much the same way that satellites dedicated to the oceans and atmosphere have vastly improved our understanding of climate and allowed us, for example, to predict El Niño events. "Atmospheric scientists demonstrated the value of spaceborne observations a little over 20 years ago, oceanographers about a decade ago, and when we have an InSAR mission, the same thing will occur for our field," she says.

In the meantime, other countries are continuing with their own InSAR projects. Japan plans to launch a ¥60-billion (US\$600-million) satellite next year, and China may launch its own as early as 2006. ■

NASA/SPR

Beta-blocker goes on trial as asthma therapy

Alison Abbott

A drug for lowering high blood pressure is to be tested as a treatment for asthma despite warnings on its packaging that it should not be prescribed to asthmatics.

The US Food and Drug Administration approved the trial on nadolol, one of a group of drugs called beta-blockers, because some animal tests have suggested it could give long-term relief from the symptoms of asthma.

The approval for the clinical trial is seen by many in the field as a vindication for pharmacologist Richard Bond of the University of Houston in Texas. Bond has been pushing for years for beta-blockers to be used to treat asthma, despite claims by colleagues that the idea is counterintuitive or even dangerous.

"Every medical student learns never to prescribe a beta-blocker to an asthmatic, because it would have a potentially fatal effect," says Clive Page, a pharmacologist and asthma expert at King's College London.

Beta-blockers treat high blood pressure by blocking receptors called beta-adrenoceptors on the membranes of cells in blood vessels. These receptors normally bind to hormones such as adrenaline, causing blood vessels to contract and raising blood pressure. Beta-blockers improve blood flow by reversing the effect.

But beta-adrenoceptors are found in other tissues, and in the lungs their activation makes the airways dilate. So drugs that block the receptors can make the airways contract, at least in the short term.



A good wheeze: Richard Bond has long backed beta-blockers as a treatment for asthma.

A class of asthma drugs called beta-agonists activate these receptors in the lungs. But a decade ago, long-term use of these drugs was linked to an increase in the number of deaths, probably because the treatment does not control the inflammation that underlies asthma. The drugs are now often prescribed in combination with anti-inflammatory steroids.

Bond rattled the cages of asthma biologists with his studies in mice, which showed that although beta-blockers initially made

asthmatic lungs more sensitive to asthma attacks, long-term treatment made them less sensitive.

Bond was inspired by a similar shift from the use of beta-agonists to treat congestive heart failure. The drugs improved heart function in the short term, but they seemed to cause more deaths in long-term treatment. Beta-blockers, on the other hand, worsened symptoms initially, but improved heart function and cut mortality later. Beta-blockers are now a routine treatment for heart disease.

Inverseon, a biotechnology company based in San Francisco, is now recruiting asthmatic patients for the clinical trial of nadolol. In the trial, the dose will be increased in weekly steps to try to avoid short-term detrimental effects.

Page says that he has been impressed with the animal data. He is now collaborating with Bond in a study testing a range of beta-blockers, including commonly prescribed drugs such as propranolol, in various animal models of asthma.

John Fozard, a veteran asthma pharmacologist with the drug company Novartis in Basel, Switzerland, points out that many treatments that have worked well in animals have disappointed in human trials. Bond acknowledges this, but says: "Even if it doesn't work in humans, the principle that compounds can have different pharmacological effects when they are given for long or short periods has been demonstrated." ■

British Museum bids to stop illicit traders using eBay

Paula Gould, London

The British Museum in London is negotiating with the Internet auction company eBay in an attempt to stem the illegal sale of archaeological finds.

The rise of Internet auction houses has made it easier to sell archaeological treasures. These are often ancient coins or pieces of jewellery, worth anything from tens to a few hundred pounds. But all such finds unearthed in England or Wales after September 1997 should, by law, be registered and viewed by experts. Even if museum officials decide not to buy the item, experts say it is crucial to record what is found.

"Some people may think that their finds are unimportant, but we still need to know about them," says Michael Lewis, deputy head of the Portable Antiquities Scheme at the British Museum.

Museum officials estimate that five potential 'treasure items' go unreported every week. A member of staff now



In the money: illegal trading in Roman artefacts, such as these coins, is rife on the Internet.

scans eBay every day for suspect sales.

Most other European countries have stricter laws governing amateur archaeology, yet these too are being flouted. In Germany, it is illegal for anyone not working on an authorized site to dig up artefacts. But ancient items are still finding their way onto eBay.

"These people know where the sites of archaeological interest are, and they are going there with metal detectors," says Wilfried Menghin, director of the Museum of Prehistory and Early History in Berlin. Staff at the Berlin museum have bought suspect relics on eBay in an attempt to trap the sellers, but have so far been unable to gather enough evidence to secure a conviction.

The British Museum wants eBay to help prevent illegal transactions, or at least to put messages on the site reminding people of their responsibilities. But eBay counters that the museum is probably better placed to keep an eye on things; the auction site's staff say they will take down any item that the museum reports to the police. ■

For more news and analysis go to
news @ nature.com
www.nature.com/news

Patent office takes small step to get the measure of nanotech

Washington The word 'nanotechnology' has been used and abused by everyone from researchers hoping to take advantage of funding to advertisers hoping to give their product a high-tech ring. But now the US Patent and Trademark Office has stepped in to provide some guidance on what nanotechnology really is.

The patent office last month unveiled Class 977 for nanotechnology patents. For an invention to qualify, the office says, at least one of its dimensions should be between 1 and 100 nanometres in size, and the tiny size of the device must be essential to its function. This means that some ingredients in sunscreen may well be nanotech, for example, but a computer chip may not be.

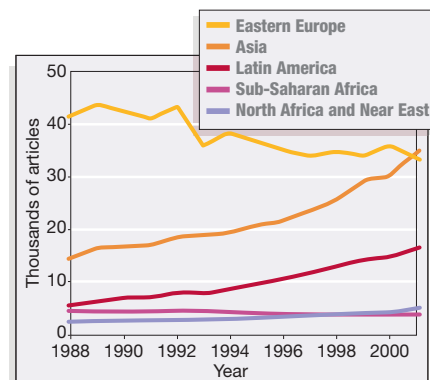
Worried about overlapping patents being granted amid a flood of claims, the patent office began training its examiners in nanotechnology concepts and terminology, and last November it set up an advisory group of outside lawyers and scientists. The new definition is meant to clear up any remaining confusion for patent officers. Whether it will have the same effect on a general public bombarded with nano-promises remains to be seen.

Latin America records rapid rise in research publications

Washington Research output is booming in Latin America and Asia, according to the US National Science Foundation (NSF).

The number of science and engineering papers published in journals with authors from Latin America tripled from 1988 to 2001, with Argentina, Brazil, Chile and Mexico leading the way. Asia is also producing more papers, mainly because of a boom in China (see *Nature* 426, 752–753; 2003).

The NSF says the rises are due to policy reform, increases in the number of researchers and the money for salaries, and



Going up: a rise in publications suggests that science is booming in Latin America and Asia.

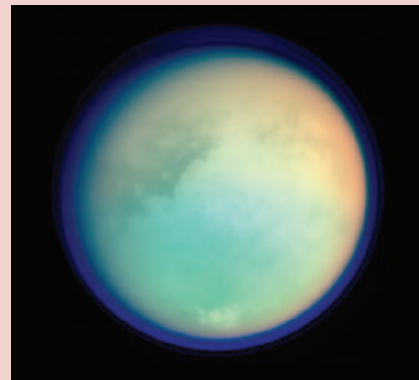
Atmospheric swoop gives Cassini a titanic view

London The Cassini-Huygens spacecraft grazed the atmosphere of Saturn's moon Titan on 26 October, taking radar images of a cold landscape carved by winds and dotted with what look like lakes.

Titan is the second-largest moon in the Solar System and the only one known to have an atmosphere. Its chemical make-up is thought to be similar to that on Earth long before life emerged, but an opaque smog makes it hard to observe Titan's surface.

The recent observations (right) indicate that Titan may have lakes of methane, ridges of minerals, and floes of ice. The surface seems to be coated in organic chemicals such as liquid propane or solid acetylene, both of which would be stable at the temperature of -179°C .

The Cassini craft is planned to swoop past



the moon 43 more times before releasing the Huygens probe in December. The probe should float through the atmosphere by parachute and gather more detailed information.

a growing tendency for researchers to collaborate in international projects.

Other regions fared less well: the research output from Eastern Europe (including the former Soviet Union) has actually fallen.

Spanish stem-cell law allows IVF embryo use

Madrid The rules governing stem-cell research in Spain have been relaxed, thanks to a bill passed on 29 October by the country's recently elected Socialist government.

Research on embryonic stem cells has technically been allowed in Spain since 2003, but only in limited circumstances and with restrictions. The creation of human embryos specifically for research was prohibited, for example, embryonic stem cells could be studied only if their use was unavoidable, and embryos left over from *in vitro* fertilization (IVF) procedures could not be used if they had been frozen for more than five years.

Some restrictions have now been lifted, provided the researchers get permission from the embryo's parents. "It's all about making things easier for scientists," says Spain's vice-president, María Teresa Fernández de la Vega.

Other European countries that currently allow stem cells to be obtained from IVF embryos include Belgium, Denmark, Finland, Greece, the Netherlands, Sweden and Britain. But Austria, Germany, France and Ireland prohibit the practice.

Caltech cash lands Earth scientists a moving study

San Diego A US\$13-million Tectonic Observatory has been established at the California Institute of Technology, launching an initiative to study the movement of the plates that make up the Earth's crust.

Seismometers, global positioning systems and satellite radar will be deployed to focus

on key tectonic plate boundaries in western North America, Central America, Sumatra and Taiwan. Their data should tackle such questions as whether tectonic plates are pushed about by new material forming at ocean ridges, or pulled by old ocean floor sliding under continents, and should provide insight into earthquakes and volcanoes.

The money comes from the Gordon and Betty Moore Foundation in San Francisco.

NASA's vomit comet sickens for a final time

Washington Wannabe-astronauts, and researchers looking for a weightless environment, will soon have a new craft in which to run their experiments.

The 'vomit comet' — the jet used by NASA to create microgravity by plunging towards Earth in a stomach-clenching dive — took its last flight on 29 October. The plane plunged 50 times during the three-hour flight, giving those on board 30-second slots for their experiments. Unusually, few of the two dozen seasoned passengers travelling on this final flight were sick.

The KC-135 jet has seen nine years of service with NASA. During this time the crew boasts to having hosted more than 2,000 students and cleaned up more than 1,000 litres of vomit.

The craft will be replaced by a C-9 plane next year that is likely to have the same sickening effect on its passengers. The first person to vomit in the new craft will probably get "a little plaque — all in good fun, of course", says test director John Yanic.

Correction

David King's Feature "The scientific impact of nations" (*Nature* 430, 311–316; 2004) states that Japan is "first on higher education" in terms of %GDP spent in G8 nations (page 313). In fact, as Figure 5 shows, Canada is first and Japan second.

Stuck on you

Recent studies are pointing the way for new uses of an ancient treatment — leeches.

Helen Pilcher wades in to find out how these creatures could help the arthritic.

They've got three jaws, with some 300 teeth, and they like to feast on human blood. "I've only been bitten around 20 times," says Carl Peters as he plunges his hand into a tank of swirling, hungry leeches. "There's about 5,000 in here," he adds casually.

Peters is assistant manager at Biopharm, a company in Wales that breeds and supplies the carnivorous worms (*Hirudo medicinalis*) for medical use worldwide. Most of the 70,000 leeches produced here annually end up draining blood from the swollen faces, limbs and digits of patients undergoing reconstructive surgery. But researchers now think that the slippery hermaphrodites may have other uses.

Recent studies suggest that leeches can lessen the pain of osteoarthritis and restore lost mobility¹. Current osteoarthritis treatments are palliative at best, so the hunt is on for any new, effective treatments — even if they involve attaching live, hungry creatures to your knees. How the beasts produce their effect is still uncertain, but scientists are keen to seek out the therapeutic proteins found in leeches, hoping then to make them synthetically and use them in the clinic.

Leeches may seem a rather outdated form of therapy, but these days they're big business. Alongside those produced in Britain, Leeches USA, based in New York, turns out thousands more and says that 10,000 are used in the United States each year. You can order your

leeches from Leeches USA for just \$7.70 each (the minimum order is seven), and have them delivered to your door, complete with a 'mobile home' for the creatures at just over \$100 a shot. And the market may be growing: in June this year, the French company Ricarimpex SAS was granted a licence to market the creatures as 'medical devices' in the United States. It was the first firm to get such a licence since regulation started in the 1970s.

Humans and leeches have a long history together. Images of leeches have been found on the walls of an ancient Egyptian tomb, and it is thought that the doctor of the Roman Emperor Marcus Aurelius used them for blood-letting some 1,500 years later². In the nineteenth century, medics thought the worms drained the body of impure blood that causes disease, and they were used to treat just about anything, from headaches to haemorrhoids.

Parasite's progress

As modern medicine developed, the tiny blood-suckers fell from favour until the 1960s, when physicians realized their potential to aid vascular surgery. During an operation, it can be hard to rejoin ruptured veins. Without functional vessels to drain blood away, tissue can be deprived of oxygen, causing cells to die. Leeches suck this excess fluid away, buying the body time to re-establish its own network of blood vessels. An astounding 80% of Britain's 62



Blood vessels: leeches attach themselves by suckers (above) and can drink ten times their own body weight in blood.

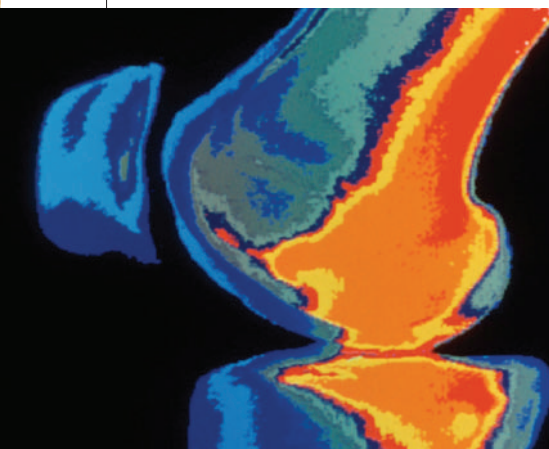
plastic-surgery units have used leeches post-operatively in the past five years².

Other uses are now being explored. Researchers Andreas Michalsen and Gustav Dobos, from the University of Duisburg-Essen in Germany, have set up clinical trials to assess the effects of leeches on knee osteoarthritis — an idea that occurred to them after speaking to local lay healers about their leech use. The degenerative condition affects up to 20% of people aged over 65, and occurs as protective tissue, called cartilage, in the knee joint breaks down, causing pain and inflammation.

In several preliminary experiments, including a trial of 24 people, the investigators slapped up to six worms on to painful joints and left them there for about an hour. The creatures drank up to ten times their own body weight in blood before they fell off. Control patients used an anti-inflammatory gel up to twice a day for a month.



Stiff drink: joints can become painful and inflamed in old age as the cartilage breaks down, but leeches ease the symptoms of osteoarthritis.



One week after treatment started, patients using leeches reported a 64% reduction in pain, as measured by self-reported questionnaire. Functional benefits lasted for up to three months. "After treatment, I could run down the stairs," says Elfriede Klein, who endured the leech therapy. "I couldn't do that before." By contrast, controls reported a 17% lessening of pain, and had no such urge to leap down staircases¹.

Since then, Michalsen and Dobos have performed a larger, as yet unpublished, study of 400 patients with osteoarthritis of the knee. A promising 80% reported a significant reduction in pain a week after treatment. After six months, about 40% said they could still feel the benefit. In the future, we may choose to reapply fresh leeches twice a year, suggests Michalsen.

Most patients had no problem getting intimate with the blood-suckers. Less than 10% of the female patients expressed disgust at their treatment, and the men said they were not repelled at all. "When they're put on you feel a gentle pinch, then a sucking feeling," says Klein. "It takes three to four minutes for them

to latch on properly, then you feel nothing at all." Around three-quarters of patients said the bite felt a little itchy, but this soon faded.

This is great when compared with the side-effects that conventional treatments can bring, points out Dobos. Non-steroidal anti-inflammatory drugs, such as ibuprofen, can ease pain sometimes, but prolonged use can lead to stomach and renal problems. Knee replacement, the other main option, involves major surgery.

But getting leeches accepted as a mainstream treatment will be hard. Large drug companies are unlikely to latch on to live creatures as a product, and may be unwilling to put derivatives from the beasts into trials until there is firmer evidence of their worth. Although the experiments show impressive results, many attribute the leech success to the placebo effect. After all, no 'proper' control could eliminate the patients' knowledge of which treatment they were using — there's no such thing as a non-blood-sucking leech. Michalsen says they have compared the patients' expectations of therapy against the actual results, and found no relationship. "The placebo effect may add to the leech effect, but it's probably not responsible for it," says Michalsen.

The secrets of saliva

There are other possible explanations for the effect: leech bites might numb a brain's response to pain, for example. Acupuncture is thought to work this way, but repetitive treatments are needed to maintain an analgesic effect. As a single leech treatment can have long-term benefits, Michalsen and Dobos think that another explanation is more likely.

Leeches don't just take, they also give, explains Dobos. As the worm bites, it injects a complex cocktail of proteins into the host through its saliva. These may be responsible for the therapeutic effects.

Certainly the saliva contains at least one anti-inflammatory molecule, called leech-derived tryptase inhibitor². Researchers suspect that the sloppy bite of a leech may contain many more such molecules.

But this may be just part of the story. Although inflammation can occur in people with osteoarthritis, researchers dispute its extent and importance. Giving patients anti-inflammatory drugs doesn't always yield improvements, and some think that any positive effects are due to the drugs' analgesic effects, rather than their ability to calm inflammation. So something else may be going on.

Osteoarthritis may be more than a disease of joint wear and tear: it may also be a vascular disorder⁴. Patients with osteoarthritis often have high cholesterol and blood that clots easily, says rheumatologist Paul Dieppe at the University of Bristol. Small clots that occur in

the blood vessels supplying bone could starve joints of oxygen and trigger the condition's symptoms, he says. Chemicals that thin the blood and break clots apart could in theory ease a patient's misery. However, Dieppe says no trials have been done with traditional anticoagulating medications for osteoarthritis: "It would be a very expensive undertaking and is somewhat speculative, so the risk-benefit issue would be a big one."

Leech saliva is full of such molecules. The best known, hirudin, was discovered more than 100 years ago⁵. The protein binds to and blocks the blood protein thrombin, a key enzyme in the clotting process. This in turn slows the production of fibrin, a tough, sticky, fibrous protein that is a key component of blood clots. A synthetic form of hirudin (lepirudin) has already made its way to the drug market as Refludan, but there could be more to follow. Another molecule, called destabilase, helps break apart any interlinked fibrin fibres; other molecules tackle the second major ingredient of blood clots: platelets⁶.

The leeches' arsenal of anti-inflammatory and anticoagulant molecules helps it go unnoticed by its host while it slurps down its liquid lunch. The beasts may also maintain a low profile by secreting a homemade analgesic. When one researcher put this to the test, she found that rats allowed to inhale a solution of powdered leech extract showed less sensitivity to pain⁷. But these results have proven hard to reproduce, and no one has been able to demonstrate a similar effect with leech saliva.

Dobos and others are now working hard to isolate and characterize the pharmacologically effective substances in leech saliva. They hope to reproduce the active ingredients chemically and to reduce the slimy treatment to a pill.

In the meantime, thousands of leeches will continue nobly to lose their lives each year in the name of medicine: once their work is done, the leeches from Biopharm and Leeches USA are gently and humanely pickled in alcohol. Life is brighter for some lucky European leeches. Those bred at the ZAUG farm in Biebertal, Germany, can expect to live out their 20 years to the full in retirement ponds. "These are hard-working animals," says Dobos. "It's only fair that they should have somewhere to grow old."

Helen Pilcher is a reporter for news@nature.com.

1. Michalsen, A. *et al.* *Ann. Intern. Med.* **139**, 724–730 (2003).
2. Whitaker, I. S., Izadi, D., Oliver, D. W., Monteath, G. & Butler, P. E. *Br. J. Plast. Surg.* **57**, 348–353 (2004).
3. Sommerhoff, C. P. *et al.* *Biol. Chem. Hoppe-Seyler* **375**, 685–694 (1994).
4. Ghosh, P. & Cheras, P. A. *Best Pract. Res. Clin. Rheumatol.* **15**, 693–709 (2001).
5. Haycraft, J. B. *Proc. R. Soc. Lond. B* **36**, 478–487 (1884).
6. Eldor, A., Orevi, M. & Rigbi, M. *Blood Rev.* **10**, 201–209 (1996).
7. Baskova, A. P., Khalil, S., Nartikova, V. F. & Pashkina, T. S. *Thromb. Res.* **67**, 721–730 (1992).

Sink or swim

Can an ambitious plan to protect unique marine habitats in the open ocean turn the tide of destruction? Henry Nicholls plunges in.

Peering through the thick glass windows of their submersible, the ocean scientists were greeted by a sight no one on Earth had ever seen before. In front of them, smoking towers of deep-sea hydrothermal vents teemed improbably with creatures, including giant tube worms and yellow mussels. As the crew of the research vessel *Alvin* stared in wonder, it probably didn't occur to them that this newly discovered biodiversity at the bottom of the ocean might one day need to be protected from people. But that was 1977, and it is now time for a reassessment.

Conservation biologists generally agree that unique habitats in the open sea such as hydrothermal vents, seamounts and cold-water reefs require urgent protection. Fishing, pollution and commercial traffic in international waters — known in treaties as the 'high seas' — have increased to such an extent that ecosystems once deemed out of human reach are feeling the effects.

Two years ago, delegates at the World Summit on Sustainable Development in Johannesburg, South Africa, agreed to establish by 2012 a network of marine reserves representing all major habitats, both within and beyond national jurisdiction. In February,

this target was incorporated into the Convention on Biological Diversity, which has so far been ratified by 188 countries. Conservation biologists are pushing for the first marine parks to be in place by 2008.

But a number of obstacles stand in the way. Uncertainty about which habitats to protect, opposition from the fishing industry, legal loopholes and the challenges of enforcing marine reserves far from land are making the targets hard to meet. Nor has there been any agreement on how, or by whom, the reserves would be run. Such hurdles are likely to dominate discussions of ocean protection at the World Conservation Congress in Bangkok, Thailand, over 17–25 November.

Much of the biodiversity in the high seas has come to light only in the past three decades. As well as hydrothermal vents, there are deep-ocean trenches, crevices that plunge to more than 10 kilometres below sea level and play host to deep-sea molluscs and scavenging crabs. Dotting the ocean floor are thousands of seamounts, extinct submarine volcanoes, each of which harbours a unique ecosystem. And some of the most abundant life in the ocean is found on cold-water reefs, delicate coral systems that grow in chilly, nutrient-rich waters.



In a bind: fishing nets can be lethal to marine mammals like this sperm whale.



Grand Banks

Once home to huge schools of cod, the Grand Banks off Newfoundland have been continually fished since the fifteenth century. Although heavily reduced, cod and flounder stocks have begun a gradual recovery since the banks were closed to most fishing in 1995. Many see the 10% of the area that lies in international waters as a good candidate for a high-seas reserve.

2

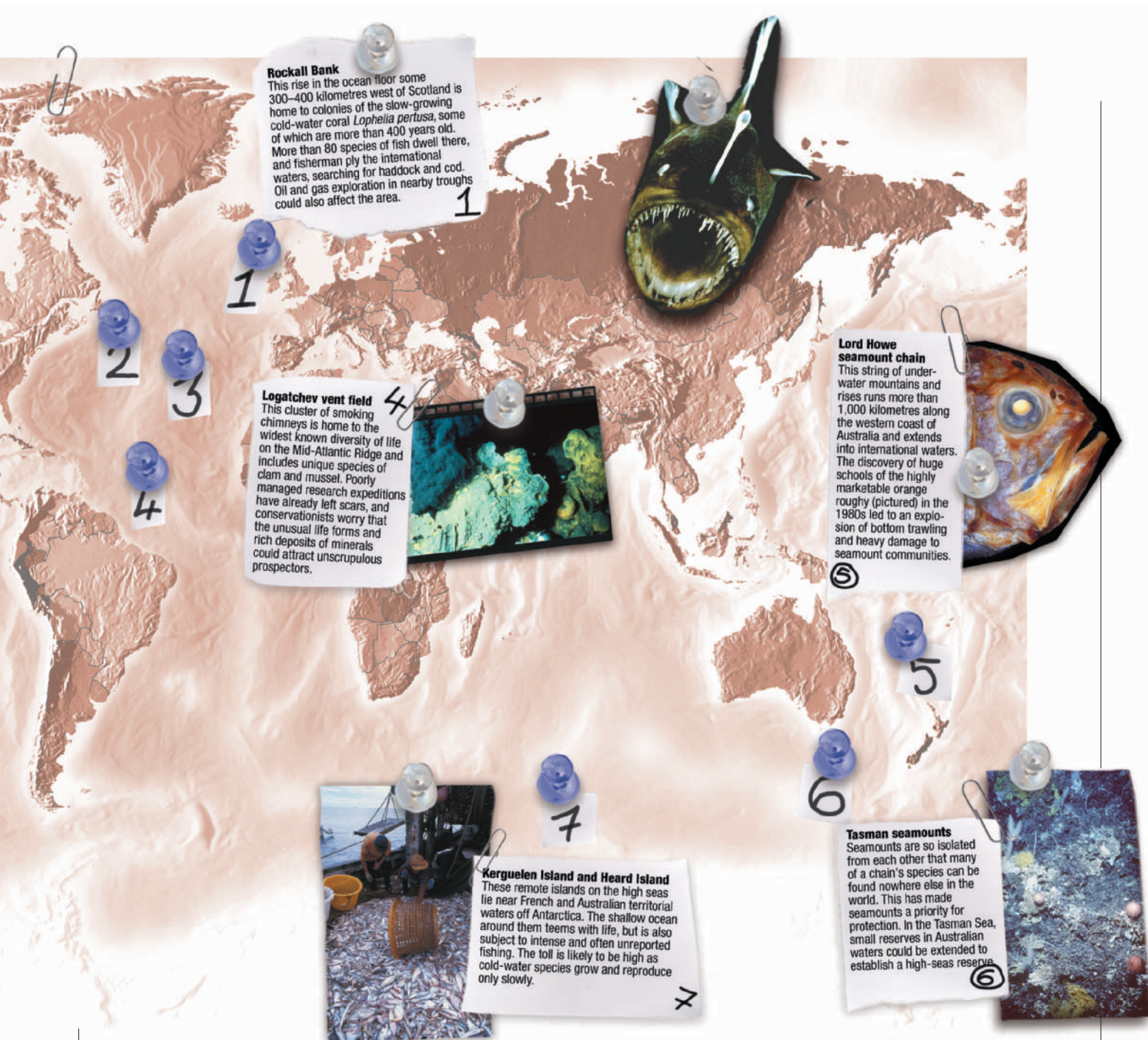
Rainbow vent field

Located just 350 kilometres southwest of the Azores, this field has more potential than other sites for exploitation by miners and bioprospectors — and even tour operators promising glimpses of deep sea creatures such as this newly discovered *Promachoteuthis*. But that proximity to land also makes it easier to police.

3

Despite decades of exploration, the species count continues to grow — each marine survey nets a menagerie of new life forms. One, conducted earlier this year by the Norwegian-led MAR-ECO expedition along the deep-sea ridges in the mid-Atlantic, turned up a new species of anglerfish and a new squid with an unusually small head and tiny eyes.

Even the open ocean — once assumed to be the marine equivalent of a desert, visited only by migrating animals such as tuna, turtles and whales — harbours a wealth of microscopic biodiversity. Last year, a single cupful of the Sargasso Sea near Bermuda was found to contain at least 1,800 new microbial species bearing more than one million previously unknown genes¹.



Park life: marine conservationists say protection is needed for sites such as these, which exemplify a range of habitats and deep-sea organisms.

"The biodiversity of the high seas is indeterminate, but it's immense," says Graeme Kelleher, who heads the high-seas task force for the World Commission on Protected Areas, the body that is coordinating protection efforts for the high-seas. Kelleher reckons that there are enough new discoveries each year to increase the estimate of biodiversity in the sea nearly tenfold. "It's not only the last great ecological frontier, but also probably the world's richest biodiversity frontier," he says.

But the high-seas habitats that are home to these organisms are being damaged. One of the main culprits is deep-sea bottom trawling, a widespread fishing practice that involves dragging weighted nets across the tops of seamounts, some 500 to 2,000 metres

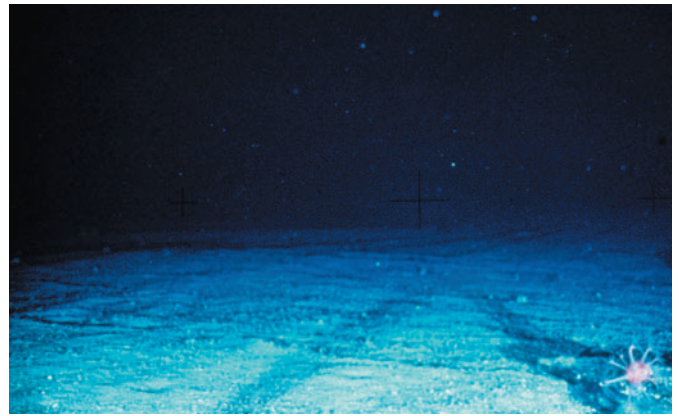
below the surface. This reduces cold-water coral reefs to rubble and decimates populations on seamounts.

In the zone

So far, marine protection has largely been limited to habitats near the coast. There are about 4,000 marine reserves within the 370-kilometre Exclusive Economic Zones (EEZs) of coastal nations, where activities such as shipping, fishing and recreational boating are restricted. In such reserves, a given country's jurisdiction provides a legal framework for protection and enforcement. Yet taken together, these reserves account for less than 0.5% of the ocean's surface. To date, there is still not a single protected area in the high seas.

Part of the difficulty is that these international waters are traditionally as open and free as the Wild West once was. That principle was made explicit in 1958 in the Convention on the High Seas, which preserved the right of all nations to exploit high-seas resources. This was superseded in 1982 by the UN Convention on the Law of the Sea, which established the conservation of marine resources as a duty of nations in principle. But without specific protection schemes, such as catch limitation, the agreement has done little to curb the intensive use of the ocean.

Although the Johannesburg meeting and the Convention on Biological Diversity set 2012 as the target date for establishing high-seas marine parks, neither specified the



Deeply damaged: fishing vessels (left) out to catch deep-sea fish such as pollack can cause serious damage to seamounts with their trawls (right).

number of areas or the expanse of ocean they should cover. Instead, the agreements called for reserves that are “representative” of all major marine habitats. This has left governments and conservation groups arguing over what the 2012 target actually means and what steps are needed to meet it.

Some guidance came from the World Parks Congress, which met last year in Durban, South Africa. There, conservation biologists agreed that at least five high-seas marine parks must be in place by 2008 if the 2012 target is to be met. Several possible sites (see graphic on pages 12–13) have been identified that could be protected by extending existing marine laws, says Kristina Gjerde, high-seas policy adviser to the World Conservation Union’s marine programme.

The legal backbone could come, for example, from the United Nations Fish Stocks Agreement, which came into force in 2001 and aims to reduce the exploitation of ‘straddling’ fish — those that cross between the high seas and EEZs — and highly migratory species such as tuna. But this is only a partial solution, as several key fishing states have yet to ratify the agreement. Furthermore, it does not contain provisions for protecting local concentrations of biodiversity that are neither straddling nor migratory, such as those living on seamounts. And in any case, fishing vessels can dodge such laws simply by registering in a country that has not signed up to them.

The right angle

Efforts to close these legal loopholes are under way, but diplomatic hurdles remain. At an informal meeting to discuss ocean issues in June, a proposal for the UN General Assembly to consider the legal framework necessary for marine parks to become a reality was thwarted by opposition from several countries, including big fishing nations such as Japan, South Korea and Iceland, says Gjerde. But if the World Conservation Union throws its weight behind the proposals for high-seas protection at the

forthcoming Bangkok meeting, the chances are that the General Assembly will address the issue this year or next, she says.

Seeking a stopgap until marine parks can be established, an umbrella group of conservation organizations called the Deep Sea Conservation Coalition is pressuring the UN to declare a moratorium on bottom trawling. On 7 October, Costa Rica formally proposed incorporating such a moratorium into the fisheries and oceans resolutions that the General Assembly passes each year. A decision is expected on 16 November.

The fishing industry is expected to fight such measures. It argues that current practices are sustainable, and that further restrictions would only harm the economies of countries that depend heavily on fishing.

Yet two-thirds of fisheries are already being fished to the fullest sustainable extent or beyond, according to UN data. And conservationists argue that fisheries actually benefit from restrictions in the long run, because marine reserves improve fish stocks both within and beyond protected areas.

Off St Lucia in the Caribbean, for instance, catches of reef fish near protected areas almost doubled within five years, says Callum Roberts, a marine conservationist at the University of York, UK². Clam fisheries in Fiji are also improving because of nearby protected areas³. “Fijian villagers now report seeing clams that are larger than they’ve seen for three generations,” Roberts says. Industrial fisheries may also benefit. For example, in 1994 a quarter of the Georges Bank near the New England coast was put off-limits to scallop dredging and trawling. Compared with the fished area, the protected area now contains about ten times as many full-sized scallops⁴.

Such benefits suggest that the costs of creating marine reserves could be offset in the long term by reducing subsidies to fishing fleets as catch numbers rise. A survey of reserves by Roberts and his colleagues suggests that a network protecting 20–30% of the ocean would cost between US\$5 billion

and US\$19 billion to run⁵. Today, subsidies to the fishing industry range between \$15 billion and \$30 billion, Roberts estimates.

But even if a legal framework can be put in place by 2012, the fisheries brought on board and funding found, there is still the problem of how to police protected areas on the high seas.

Great scheme

Technology may provide part of the answer. On Australia’s Great Barrier Reef — widely recognized as a model marine protected area — all boats must be equipped with satellite navigation systems and transmitters. These devices tell skippers which of six different zones they are sailing through and hence what activities are permitted. They simultaneously transmit to park authorities the precise location of the vessel.

The system has performed well, even under duress, says Kelleher. At the beginning of July, the ‘no-take zone’ — where all fishing is prohibited — expanded from less than 5% to a third of the protected area with no significant problems. But there is little doubt that such a system would be much harder to operate on the high seas, not least because a larger area much farther from shore would have to be patrolled to spot any vessels that were not fitted with transmitters.

With so many complicating factors, few are willing to make a firm prediction on whether the 2012 target can be met. Daniel Pauly, a leading fisheries biologist and director of the Fisheries Centre at the University of British Columbia in Vancouver, Canada, says that the biggest factor is likely to be how much pressure conservation groups can exert on the UN and on individual nations. “It will depend on political will,” he says.

Henry Nicholls is a freelance writer based in London.

1. Venter, J. C. *et al.* *Science* **304**, 66–74 (2004).
2. Roberts, C. M., Bohnsack, J. A., Gell, F., Hawkins, J. P. & Goodridge, R. *Science* **294**, 1920–1923 (2001).
3. Gell, F. R. & Roberts, C. M. *Trends Ecol. Evol.* **18**, 448–455 (2003).
4. Murawski, S. A. *et al.* *Bull. Mar. Sci.* **66**, 775–798 (2000).
5. Balmford, A., Gravestock, P., Hockley, N., McClean, C. J. & Roberts, C. M. *Proc. Natl Acad. Sci. USA* **101**, 9694–9697 (2004).

European council should back young investigators

Support for junior researchers would be fair to all countries and disciplines.

Sir — Robert May, in his Commentary article “Raising Europe’s game” (*Nature* **430**, 831–832; 2004) rightly argues for autonomy and scientific leadership for the proposed European Research Council (ERC), as well as emphasizing the need to promote scientific excellence and to foster young scientists. There is broad consensus on these priorities. One could add the need for grants instead of the contracts that European Commission (EC) programmes currently use. Research is an evolutionary and creative process that requires flexibility; but with a contract, both the project and its outcome are fixed at the outset and the money can only be used according to these terms, except after uncertain, time-consuming renegotiation.

There is an urgent need to get the ERC up and running — preferably by January 2007 to coincide with the launch of EC Framework Programme 7. I suggest that this can be achieved by focusing initially on young investigators.

May argues cogently for a focus on excellence in basic research and a lean administration, using as comparison the US National Science Foundation (NSF). The aims are valid, but the comparison is

not entirely appropriate. Each European Union (EU) country has its own agencies for developing science policies and funding research. The ERC, with a potential budget of €1 billion (US\$1.26 billion), is not meant to compete with or replace these bodies, but to bring a new European dimension. The scale and structure of the NSF (40 divisions, 250 programmes and 1,500 staff on a budget of \$5.5 billion) cannot be replicated in Europe, as it will require much closer European integration at all levels.

The success, or otherwise, of the ERC will depend on satisfactory responses to two major challenges: political endorsement and scientific acceptance. Politicians in certain countries fear the absence of *juste retour*: a share of the funding proportional to their contribution. This fear is exacerbated by concerns that their scientists will struggle to compete successfully in an ERC based exclusively on scientific excellence. Scientists also worry that the ERC will suffer from the same oversubscription problems as do existing EU programmes.

There is one way forward that could address these concerns, help keep the ERC manageable and be acceptable to many.

I recommend that the initial programmes of the ERC should focus predominantly, if not exclusively, on supporting young investigators who are establishing their first independent group. These programmes should go beyond existing European initiatives, such as the young investigator awards distributed by the European Heads of Research Councils, or the Wellcome Trust career development schemes.

This approach would meet with approval from the scientific community in most, if not all, disciplines. It would put all countries on an equal footing, for they all have excellent young researchers. It would be easier and quicker to implement than programmes with defined research priority areas, and should reduce the problem of oversubscription by reducing the potential customer base. And it would have a real impact on the European science base. This could help to define the ERC’s visibility, usefulness and long-term prospects, laying the foundation for further development.

Luc Van Dyck

*Initiative for Science in Europe (ISE),
c/o European Life Sciences Forum,
Meyerhofstrasse 1, D-69117 Heidelberg, Germany*

US rules on tech transfer to foreign nationals

Sir — I would like to respond to your News story “US universities up in arms over licence plans for foreign staff” (*Nature* **431**, 615; 2004), for which I was interviewed. Although the US Department of Commerce appreciates *Nature*’s coverage of this important issue, it is necessary to correct certain inaccuracies regarding US export control policy for technology transfer to foreign nationals.

First, the article says that a “change ... is being proposed by the [commerce] department”. The department’s independent inspector-general has indeed recommended certain revisions, but the department itself has not proposed any change.

Second, the story states: “Historically, universities and government laboratories doing basic research have been exempt from these rules.” This is incorrect. Although there are exceptions for some technology transfers by the research community (such as for ‘educational information’ or for technology that arises during fundamental research), the long-standing regulations have never provided a blanket exception for the research community from the

technology-transfer rules.

Third, the article repeatedly suggests that a licence would be required merely because a foreign national operates equipment that the United States controls for export. This too is incorrect. It is the transfer of controlled ‘use’ technology to foreign nationals (for example, training someone in how to use controlled equipment) that may require a licence, not merely the use of such equipment. In other words, there is a critical distinction — absent from your article — between the transfer of controlled ‘use’ technology, and the use of controlled equipment, which does not require a licence.

Finally, the article incorrectly attributes to me the statement that “a licence would be required if technologically significant information was gained from using the equipment”. The point I made was that only certain ‘use’ technology is sensitive: not all technology for using controlled equipment is controlled. Moreover, assuming that sensitive information was gained from using equipment, the information would not be controlled if it resulted from fundamental research.

Given the complex public-policy issues involved here, it is crucial for all stakeholders to have an accurate understanding of the situation (see also

www.bis.doc.gov/DeemedExports/DeemedExportsFAQs.html).

Peter Lichtenbaum

*Assistant secretary for export administration,
US Department of Commerce, Washington DC
20230, USA*

Industry funding doesn’t influence our reports

Sir — U. Rangan and J. Shacter of Consumers Union, in Correspondence (*Nature* **431**, 505; 2004), comment on our industry funding. Members of our Scientific Advisory Panel, which at present consists of more than 350 leading scientists and physicians, independently review all our scientific publications. These reviewers, of whom Joseph Rosen is one, serve on our advisory board and critique our work without compensation, just as reviewers of publications such as *Nature* do. We have never allowed, and will never allow, a funding body to influence our scientific positions. Interested parties can access our publications at www.acsh.org.

Ruth Kava

*American Council on Science and Health,
2nd floor, 1995 Broadway, New York,
New York 10023-5860, USA*

Revolutionary change

The French Revolution had a profound effect on the nation's science.

Science and Polity in France: The Revolutionary and Napoleonic Years

by Charles Coulston Gillispie
Princeton University Press: 2004. 664 pp.
\$80, £51.95

Bruno Belhoste and
Bernadette Bensaude-Vincent

At the end of the eighteenth century, while England was experiencing the Industrial Revolution, France was going through two simultaneous revolutions. The political revolution changed a rigid order into chaos, and later into tyranny, with a rapid turnover of regimes taking it from a kingdom to a republic and empire. The scientific revolution, quietly conducted by a host of mathematicians, astronomers, chemists and physiologists, brought about new theories in mathematical physics, chemistry and biology, as well as a new regime of knowledge production with the creation of higher-education institutions.

This period of French scientific life is an ideal case for those who want to study the subtle interplays of science and politics. The US historian Charles Coulston Gillispie, who founded the history and philosophy of science programme at Princeton University in the 1960s, has devoted most of his scholarship to the study of French science around 1800. This book continues the broad sweep of work he initiated in *Science and Polity in France: the End of the Old Regime* (Princeton University Press, 1980). The first book presented a 'golden age' when France was at the centre of European scientific life, and analysed the cultural, political and technical factors that prompted this leadership. This sequel doesn't just describe the many changes and innovations that occurred in various scientific fields at the turn of the nineteenth century, but disentangles the intricate threads that linked scientists and politicians during the revolutionary and Napoleonic years.

The French Revolution brought scientists into power. The mathematicians Lazare Carnot and Gaspard Monge, and the chemist Louis-Bernard Guyton de Morveau, used their analytical skills to organize the military defence of the country. The Marquis de Condorcet, a mathematician and secretary of the Academy of Sciences, was in charge of reforming the educational system as a member of the Legislative Assembly and the Convention.

The crossover between science and politics went both ways. During his emergence as a political and military leader, Napoleon

Bonaparte flirted with science, solving a geometrical problem to divide a circle into four equal parts using just compasses. Even at the height of his career he enjoyed the company of scientists, and his expedition to Egypt was conceived as both a military conquest and a scientific expedition, involving astronomy, zoology and chemistry.

As in his previous book, Gillispie argues here that the close alliance between knowledge and power was profitable for both sides. Savants were able to secure the financial and institutional support necessary for the advancement of knowledge. In turn they provided politicians with instruments for increasing their military power, and experts helped them to secure social control.

More fundamentally, the French Revolution brought science and civil society closer. In the old regime, individual roles were defined by social status, under a system of estates and orders. The new republic of science introduced a meritocratic value system that was strikingly different. Similarly, in the modern society that emerged from the political revolutionary upheavals, social roles were redefined according to function, and the notion of merit became the main regulator for access to positions in the state, much as in academic life.

Gillispie suggests an interesting parallel between this societal change and a simultaneous shift in science. The eighteenth-century encyclopaedic culture of natural philosophy was overthrown by a disciplinary reorganization of science with the creation

of new educational institutions. The traditional quest for truth, with its emphasis on taxonomy, was discredited by the positivist call for scientific enquiry to be confined to describing phenomena and establishing the natural laws that rule them. "In both politics and science the premium is on effectiveness, on doing something rather than being someone," writes Gillispie. He is content to point out the analogy between epistemological and political choices without venturing any interpretation. He recognizes in Auguste Comte's *Cours de Philosophie Positive* an image of the new epistemic regime that emerged, but does not follow Comte's views of the links between science and politics.

Such a cautionary attitude exemplifies Gillispie's historiographic style. His method is based on sound empirical data and he is wary of grand philosophical claims. He thinks that most historians who study this period place too much emphasis on theoretical and ideological debates. In his view, the deep and lasting changes concerned the daily practice and conduct of science, rather than world views.

Gillispie is a fair philosopher of science but a great master of erudition and historical narratives. This volume can be read as either a saga of science or a series of short, loosely interconnected stories. The author is at his best when he is reporting colourful episodes, portraying a character, disentangling a plot or dissecting an institution. Although this book is intended for a scholarly audience of science historians, it will also appeal to



Voyage of discovery: Napoleon's journey to Egypt was a scientific trip as well as a military conquest.

THE ART ARCHIVE/MUSEE DU LOUVRE, PARIS/DAGLI ORTI

scientists who are fond of history. They will particularly enjoy the chapter on the creation of the metric system, the narrative on Condorcet's death, the subtle analysis of the Monge connection, and, above all, Napoleon's fascinating and exotic expedition to Egypt. ■

Bruno Belhoste and Bernadette Bensaude-Vincent are professors of the history of science, Université Paris X, 200 Avenue de la république, 92001 Nanterre, France.

A brain in the hand

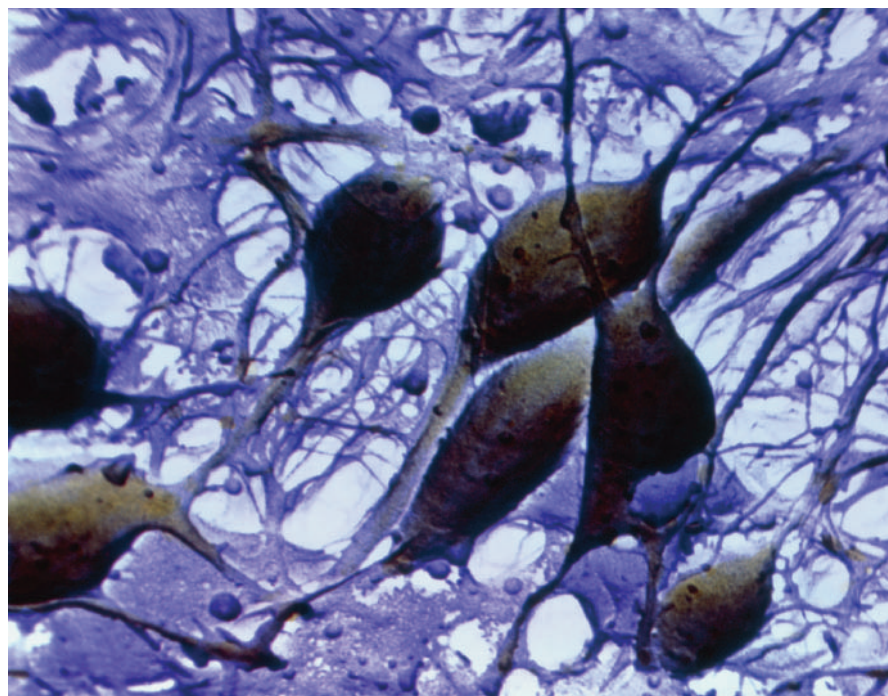
On Intelligence: How A New Understanding of The Brain Will Lead to the Creation of Truly Intelligent Machines

Jeff Hawkins, with Sandra Blakeslee
Times Books: 2004. 272 pp. \$25

Igor Aleksander

With so many books published recently on minds, machines and intelligence, it is becoming progressively more difficult to find their distinguishing features. Happily, Jeff Hawkins has a unique perspective, as one of the pioneers of hand-held electronic organizers and founder of the Palm Computing and Handspring companies, whose products adorn the pockets of executives and technical boffins alike. His hard-nosed grasp of complex digital-system design and his passion for neuroscience are the authoritative bases for this book.

On Intelligence may not herald a future of super-intelligent hand-held computers, but it does analyse how this objective could be approached from an understanding of some particular and complex computational features of the brain. It is written as a personal history, with step one for Hawkins being the realization in the mid-1980s that computers and research into artificial intelligence had not produced the kind of intelligence that each and every one of us possesses. He has always felt that artificial intelligence, being based on achieving smart outward behaviour through the speed and power of a computer, tells us nothing about what intelligence is, and why brains are better at it than computers. He argues that the brain can be understood only if its intricate architecture can be analysed and turned into computing theories. His ideas were rejected by both industry and the academic world at a time when the stranglehold of the 'power and speed' approaches of the 'artificial intelligentsia' did not recognize the power of brain models, caricaturing the brain as 'slow and squishy'. Well into his career as a successful computer engineer, Hawkins studied biology and neurology at night and then gave up his day job to become a graduate student in biophysics at the University of California, Berkeley.



Smart thinking: understanding the cerebral cortex can aid the development of intelligent computers.

Hawkins made the move just as neural networks, or connectionism, became the favoured alternative to classical artificial intelligence. But this too proved to be a disappointment. Neural networks were only distantly inspired by the brain and this remoteness was fatal. Connectionism became obsessed with the mathematics of learning and revealed little about the nature of real intelligence. The brain, as the most complex machine on the planet, remained unexplored in the computational sense.

Depressed by these failures and the seeming flood of uncoordinated data that neuroscience produces, Hawkins started looking for organizing principles. He turned to the work of Vernon Mountcastle of Johns Hopkins University in Baltimore, who advocates looking for common mechanisms in different modalities. Although vision and hearing are different, for example, the ways in which signals are processed in the two cases bear an engineering similarity that points to general principles underlying both.

Much of *On Intelligence* could be summed up as 'how computer scientists got it wrong'. Memory in the brain is an active affair that stems from the interaction of cells. It has stable states that provide the 'Aha!' sensation when an input trigger causes part of the brain to fall into such a state. This is a long way from the filing-cabinet type of memory in conventional computers. The brain is constantly trying to predict its input, whereas a computer merely waits for the input and reacts to it. These differences lead to clear expositions by Hawkins of the way the cortex may be organized to achieve this active engagement with the world, which is true intelligence. The working of the cortex

leads Hawkins to comment on the fallacy of treating consciousness "like a sauce" that turns meat into a conscious being; instead he makes a series of testable predictions about the power of his memory-prediction model.

Anyone looking for a blueprint for the ultimate conscious, palm-held computer will be disappointed. This is not a book on how to compute, it's more about how not to compute. The book is clear and punchy, and is fine for general reading, thanks in some measure to Hawkins' co-author, Sandra Blakeslee, a columnist at *The New York Times* who has helped several authors improve their books about mind and psychology. As with other computational experts who have written about the mind, Hawkins could be criticized for not having noticed that many computer scientists share his views. The time has come for people who share these ideas to stop deprecating previous efforts and look at each other's work. Other authors have suggested architectures like those proposed by Mountcastle (without mentioning him), for example, and analyses of dynamic memories and prediction have been appearing in the literature since the mid-1990s.

Nevertheless, Hawkins makes an appealing case for the sort of computational analysis that is likely both to clarify how the brain works and to give artificial-intelligence systems a new grounding. Such work requires more collaboration among those who agree with Hawkins. The fact that two years ago he created the Redwood Neuroscience Institute should help. ■

Igor Aleksander is emeritus professor of neural systems engineering at Imperial College London, Exhibition Road, London SW7 2BT, UK. His book *The World In My Mind* will be published in 2005.

Playing host to evolution

Infectious Disease and Host-Pathogen Evolution

edited by Krishna R. Dronamraju
Cambridge University Press: 2004. 384 pp.
£65, \$95

Sunetra Gupta

It is apt that in the year that marks the 40th anniversary of the death of J. B. S. Haldane, we should have a book on host-pathogen evolution that so explicitly acknowledges the debt this field owes to him. I refer not just to his ideas regarding the influence of infectious disease on host evolution, but also to his tireless efforts to inject rigour into the language used to describe it. "Words," he said, "are well adapted for description and the arousing of emotion, but for many kinds of precise thought other symbols are much better." His genius lay not only in defining a set of problems that would keep so many of us gainfully employed, but also in identifying some of the tools with which to try and tackle them.

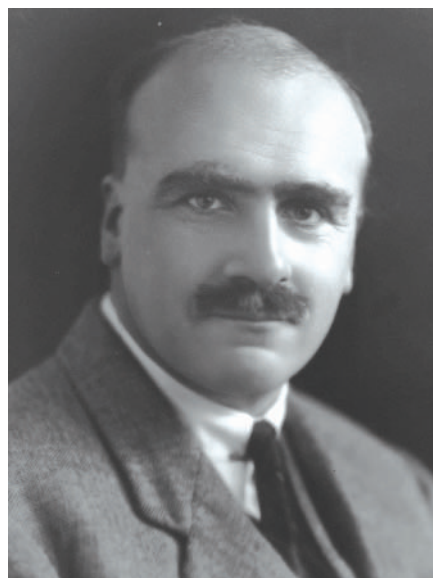
The book opens with two wonderful essays. The first, by James Crow, locates Haldane's ideas on disease and evolution within his multiple and diverse intellectual occupations. The second, by David Weatherall, follows the fascinating trajectory of Haldane's 'malaria hypothesis' — the proposal that the high frequency of the blood disorder thalassaemia could be attributed to selection by the malaria parasite *Plasmodium falciparum*. "It is doubtful," Weatherall comments, "whether Haldane could have had any idea of the large industry that his short paper of 1948 had spawned." So, just how well does this book exploit this?

Sara Tishkoff and Brian Verelli's chapter on glucose 6-phosphate dehydrogenase deficiency and malarial resistance in humans, and Peter Zimmerman's discussion on the link between the Duffy blood-group antigens and resistance to *Plasmodium vivax* malaria, fit neatly into the scheme. The chapter by Alain Dessein and others on the genetics of host resistance to *Leishmania* and schistosomes carries the theme into the realm of other parasites; Tom Little and Dieter Ebert's summary of their elegant work on *Daphnia* and its diminutive parasites broadens the context to include a different host. The latter is a model chapter, and refers intelligently throughout to the evolutionary principles articulated by Haldane. Extending some of these principles to public-health concerns, the chapter by Andrew Read and co-authors, which outlines with due caution their theory that vaccines may increase virulence, also finds a rational place in this collection.

Newton Morton's rather sketchy chapter on genetic epidemiology provides a useful comment on both its history and its methodology, but might have more properly belonged in the introduction. So too would Michel Tibayrenc's endorsement of the Human Genome Diversity Project as a means of answering certain critical questions surrounding susceptibility to infectious disease. Finally, extending haldanesque thinking to diseases of unknown origin, the chapter on diabetes by Kyle and Gregory Cochran is a refreshing addition to a book that could have restricted itself to diseases that are obviously infectious.

But set among these are several chapters on issues that do not easily integrate into the general theme of Haldane's legacy. Two excellent, detailed chapters on the evolution of plasmodia could claim to have come in on the malaria ticket, but those on the evolution of influenza and cholera, although of high quality, do not find their natural home here. More inexplicable is the inclusion of a chapter on regulatory DNA.

There is a good reason to be so concerned about theme. The chapters are based on peer-reviewed work that has appeared in high-quality journals, and the authors have, almost without exception, devoted time and effort to put their findings together in a clear and intelligible manner. But who will read this book? Long gone is the age where "definitions first appeared in books rather than articles", to quote from Morton's chapter. The purpose of putting together such a book is surely to inspire students and teachers of biology at an advanced level, and thematic unity is a critical element in achieving this goal. In this respect, the book's lack of continuity is detrimental to the ever-diminishing dialogue between scientists and eager students of science. This gap is unlikely to be bridged



J. B. S. Haldane laid the foundations for studies of how infectious disease affects host evolution.

by expressions of embarrassment at the "Eskimo" custom of lending their wives to dinner guests (in the concluding chapter by Luca Cavalli-Sforza), even if the reference is to the *Lonely Planet* guidebooks. One can imagine such a gross example of the commoditization of women appearing in one of Haldane's essays for the *Daily Worker*, but it is doubtful that it would have found its way to the part of his mind that dwelt upon disease and host evolution.

Sunetra Gupta is in the Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.

Populations in space and time

Ecology, Genetics, and Evolution of Metapopulations

edited by Ilkka Hanski & Oscar Gaggiotti
Academic Press: 2004. 696 pp. \$54.95, £36.99

Alan Hastings

Edited compilations of papers can often be the only place to find an integrated and timely view of a rapidly developing subject area. This volume on metapopulation dynamics is the third to be co-edited by Ilkka Hanski in the past 15 years. Given the editors' decision to commission new chapters, rather than to revise the previous books, it is perhaps best to view this one not as a fully developed, integrated view of metapopulation dynamics, but rather as the latest, extended issue of a journal — perhaps entitled 'Trends in Metapopulation Research'.

The concept of metapopulations is one way to include the role of space in population dynamics. The original models for metapopulations focused on the fraction of occupied patches or habitat as the measure of population size, and included only colonization and extinction as processes, but more recent work has become increasingly sophisticated and general. Any treatment of the subject needs to ask whether biological questions are being forced into the metapopulation framework of local colonization and extinction when other approaches to spatial dynamics might be more appropriate, and whether the metapopulation approach should be extended — is it more than a way to understand butterflies?

The book's overview chapters focus on connecting metapopulation research to other advances in studying spatial ecology, including continuous-space stochastic models and landscape approaches, but in my view this section could have been extended. Some connections, such as those to recent work on ideas of scaling in spatial ecology, are not discussed in enough depth.

Much of the current ecological work on

metapopulations can be traced back to work by Richard Levins in 1970, but similar kinds of issue were raised in a genetic context by Sewall Wright as long ago as the 1930s. Also in the 1930s, A. J. Nicholson and V. A. Bailey, along with G. F. Gause, raised ecological questions in their classic contributions, but without using the term 'metapopulation'. Yet, as discussed in the introductory chapter, research in this area has increased sharply in the past 15 years, making any attempt at an overview difficult.

In contrast to the earlier volumes co-edited by Hanski, this one gives almost equal

weight to the evolutionary and ecological aspects of the subject. There are also fewer overviews of metapopulations, so this is not a volume to teach newcomers the basics. Instead the emphasis is on more specialized topics and applications, including conservation and reserve design, disease and pathogen dynamics, and speciation.

The papers in this volume use primarily theoretical and conceptual approaches. Readers looking for data-heavy contributions to help place the conceptual work in context will be disappointed. Also, although both genetic-evolutionary and ecological

questions are considered, few contributions truly synthesize these different disciplines.

Not many readers are likely to read this massive volume from cover to cover, but it should prove useful as a reference book. There is good coverage of enough important recent advances, such as the use of coalescents and the study of metacommunities, so at least parts of the book will be required reading for students and researchers in spatial ecology. ■

Alan Hastings is in the Department of Environmental Science and Policy, University of California, Davis, California 95616, USA.

Science in culture

Facial diversity

An exhibition in London features the changing expression and representation of the face.

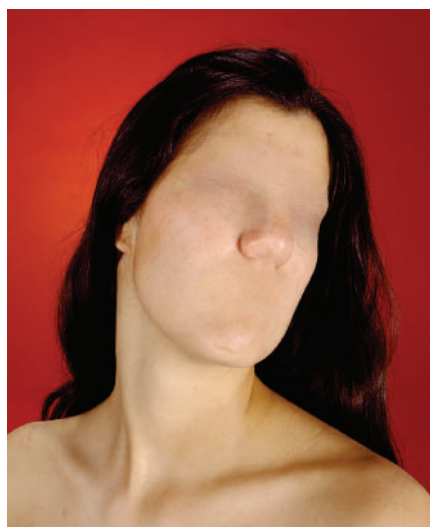
Jonathan Cole

The face is both a unique identifier and an embodied read-out of emotion. We look through it to see — or think we see — personality. And we use the face as a means of expression, to communicate with others, with and without words. Yet we now see so many unknown faces, in the street and in the media, that sometimes we gaze on others with a detached curiosity that can never be satisfied, seeing but not engaging. Sandra Kemp is well aware of this; her wonderful exhibition, which is presented by the Wellcome Trust and can be seen at London's Science Museum until 13 February, invites us to look and admire faces but also, constantly, to question and to feel.

In the first room, a series of scientific representations explore where the face begins and ends. Contemporary images from scans of faces and of blood vessels in the face and head combine accuracy with beauty. An 1837 drawing of a dissection has presence and poise. Opposite these are ancient, exquisite sculptures of faces, presaging Picasso and Modigliani. The scientific and the artistic complement each other in content and aesthetics throughout.

At the heart of the room is a collection of masks, used to conceal and transform identity because of disease, in Japanese Noh theatre, or to hide the face of an executioner. Beyond this are pictures of facial injuries from the First World War and the fitting of prostheses for camouflage. Across the room, a middle-aged man in a tweed jacket has his enucleated eye socket dressed.

A video installation by Chris Dorley-Brown has two screens. One screen shows a series of 10-year-olds who were given 15 seconds to express themselves in their faces; the other screen



shows the same people doing the same a decade later. At the age of 10 they were animated, alive and playful, but by 20 they were concerned, knowing and apparently less, or at least differently, embodied in their faces. There is also a short series of people with facial disfigurements,

each enjoying life and wearing a big smile.

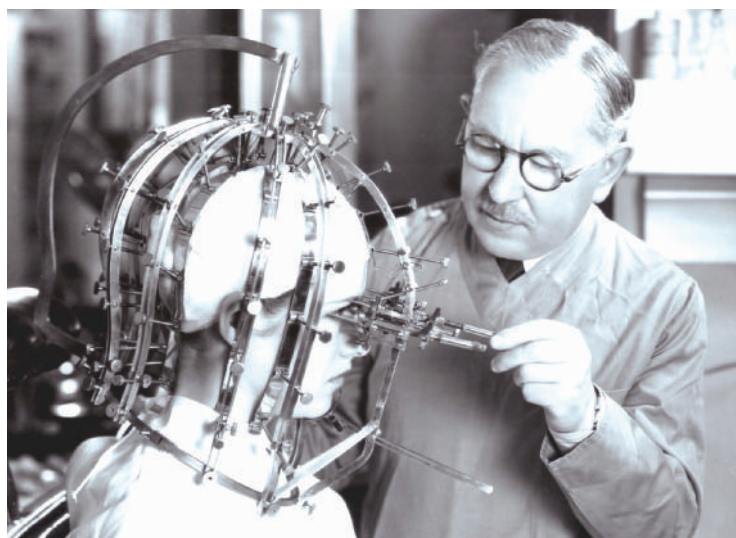
Max Factor (the man behind the brand) appears trying in vain to measure facial imperfections with steel gadgetry, as shown below. Elsewhere, a tattooed face shows too much, and a Geisha reveals little. Hollywood's monsters lurk in a series of small photos. Dolls lie in glass cases; one has several faces, seen by rotating the head, the posh face happy, the poor miserable. A computer gives the next doll facial expressions and responsive emotions — but is she comforting or scary?

We are given a view of recent developments in the recognition of facial features and expression before we see examples of how new technology can alter our perception of the face. We meet Kaya, a virtual model, given imperfections to improve her humanity, and an avatar who talks to us face to face, a presence from a strange virtual world. Then we are seduced by a digital Marlene Dietrich.

Kemp shows us that we have always portrayed faces in culture, often distorting them, but that, through new technology, we will do this even more powerfully in the future. Might this alter our views on facial attractiveness, or even alter the semiotics of facial expression? This small gem of an exhibition

delights and questions throughout. It is complemented by an excellent book, *Future Face* by Sandra Kemp (Profile Books, £12.99), and an interactive CD.

On the way out we pass a gleaming black-and-white photo of a man yawning, an image that is almost unnerving and completely contemporary. Extraordinarily, the original was sculpted in the eighteenth century. Where, Kemp asks, will our imagination and new technology take facial representation in the future? ■ *Jonathan Cole is at the University of Bournemouth and is a clinical neurophysiologist at Poole Hospital, Longfleet Road, Poole BH15 2JB, UK.*



Making sense

Proprioception: is the sensory system that supports body posture and movement also the root of our understanding of physical laws?

Victor Smetacek and
Franz Mechsner

Aristotle argued that human beings have five senses at their disposal. Although various other sense organs have come to light since then, this antique dogma still constrains popular imagination. The term 'sixth sense' resonates with instinct and metaphysics, implying that although the five 'regular' senses represent reality adequately, there is something else lurking in the subconscious. The search for the sensory system with which the blind guide their movements revealed that the body's sense of posture and movement relies on different types of tiny receptors densely packed in the muscles and tendons. In 1906 Charles Sherrington coined the term proprioception (perception of one's own) for the sensory modality based on these receptors and called it our 'secret sixth sense'. But this concept of the body as a major sense organ has failed to arouse the interest it deserves.

Proprioception functions in much the same way as the conventional senses. Proprioceptors precisely measure physical properties, such as muscle length, tendon tension, joint angle or deep pressure. Signals from this sensory orchestra are sent by afferent nerves through the spinal cord to the somatosensory, motor and parietal cortices of the brain, where they continuously feed and update dynamic sensory-motor maps of the body. So proprioception provides information on the physics of the body, the momentary distribution and dynamics of masses, forces acting on the limbs and their highly nonlinear interactions. The maps derived from these complex calculations not only guide body movement, they also (together with touch) sense the size and shape of objects and measure the geometry of external space. Weight — one's own and that of objects — is measured independently by pressure sensors and muscular tension. So subjective body consciousness provided by myriad networking proprioceptors is the basis of objective knowledge of fundamental physical properties — space, time and weight — of external reality.

Our daily doings are coordinated and run by a trinity of independent sensory systems: proprioception, vision and the vestibular organs of the inner ear (which sense balance, momentum and guide the eyes). Their signals are so tightly integrated that it is impossible to unravel them through introspection, a view which seems to favour vision



Effortless grace? A body's fine-tuned actions belie inner complexities.

as the primary sense organ of the mind. But whereas in the congenitally blind other senses more or less compensate for the loss, a child born without proprioception would not know it had a body and would be physically and mentally retarded as a result.

Selective loss of proprioception in adults is rare. In the case of Ian Waterman, a rare disease caused degeneration of sensory nerves relaying information from the body to the brain from the neck down, but spared the motor nerves conveying signals in the other direction. He could see, but not feel, where his body was or whether it was moving or not. At the age of 19, he was left a helpless 'rag doll', who had to be fed, washed and dressed — attempts at movement elicited only uncontrolled jerks. However, his strong will and memory of his body enabled him to learn to gradually control and guide his movements with his eyes. But even after 30 years of intense practice, the simplest movement has not been automated, but requires concentrated visual attention so strenuous that he likens it to a daily marathon, and in the dark he still collapses like a rag doll. His case, and a few others, demonstrate that all purposeful movements, both conscious and unconscious, are controlled by proprioception.

The proprioceptive system is so efficient and reliable at granting us the freedom of movement we expect from our bodies that we unconsciously relegate it to a subconscious, background realm of reflexes below the sphere of the five 'primary' senses. This attitude is unjustified. Most of our movements are indeed automated and run, as in animals, by the more basic and evolutionarily older parts of the brain. But we easily forget, for good reason, the intense conscious attention required to learn complex skills, such as

writing, skiing or driving. Learning a skill implies developing new patterns of movement by screening, coordinating and calibrating relevant information from the orchestra of signals supplied to the neocortex by the trinity of sensory systems. New neural programmes are computed, memorized by repetition and transferred to the more fundamental regions of the brain, from where they are run with less effort and relayed much faster than from the seat of

conscious awareness. So mastering a skill amounts to automating it.

Human proprioceptive ability is far superior to that of animals, reflected in the range, variety and precision of our automated skills, and manifest in the tools we make and what we achieve with them. Just as tools are extensions of the body, so basic scientific instruments — the balance, pendulum and measuring rod — are extensions of body sense. Evolution of these instruments, together with optical ones, launched the scientific exploration of space and time domains outside those of the body experience. Interestingly, the models of external space based on mathematical language made sense before their confirmation by precise measurements. From where else could these models have come, if not from the neural correlates of internal models of the physical laws of motion which run our bodies automatically? So the rules and laws of science were in place, and obeyed blindly, before they were rediscovered in the external world. In short, the neural correlates of physics and mathematics did not evolve *de novo*, but are rooted in our 'subconscious' body sense. ■

Victor Smetacek is at the Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, D-27570 Bremerhaven, Germany.

Franz Mechsner is at the Institute for Occupational Physiology, Ardeystr. 67, D44139 Dortmund, and the Institute for Advanced Study, Delmenhorst, Germany.

FURTHER READING

Angelaki, D. A. *et al.* *Nature* **430**, 560–564 (2004).
Cole J. *Pride and a Daily Marathon* (MIT Press, 1995).
Sherrington, C. S. *The Integrative Action of the Nervous System* (C Scribner's Sons, 1906).
Smetacek, V. *Nature* **415**, 481 (2002).

Wandering nostrils

Philippe Janvier

A small matter of head anatomy has long been a cause of controversy among those interested in vertebrate evolution. An answer that may prove generally palatable now emerges from an ancient fossil fish.

The structures known as choanae may seem obscure. But we've all got them; they are the 'internal nostrils' that form the passage between our nasal cavity and throat that we use for breathing when our mouth is closed. They have also been the subject of much argument among those studying comparative vertebrate anatomy — in particular, the question of how choanae originated in the tetrapods, or land vertebrates, a group that consists of amphibians, reptiles, birds and mammals.

This unique feature of the tetrapod palate was first regarded as an adaptation to breathing air. Now, however, we know

that choanae occurred first in extinct fish relatives of the tetrapods at least 380 million years ago — that is, 30 million years before tetrapods developed limbs and took to the land. So choanae may not have been initially involved in breathing. But did they arise as a novelty within the ancient fish relatives of tetrapods, or were they derived from a pre-existing structure in other fishes? This is a debate that has lasted for about a century, but it is practically settled by new data presented on page 94 of this issue by Zhu and Ahlberg¹. The subject of their studies is *Kenichthys*, a 395-million-year-old fossil fish from China. The background to the story,

described below, is also outlined in Figure 1.

Most living jawed fishes have two external nostrils on each side of the snout: an anterior one allowing entry of water into the nasal cavity, and a posterior one for its exit. In contrast, tetrapods have only one pair of external nostrils, but their nasal cavity communicates with the mouth cavity by means of the choanae. A single group of living fish, lungfishes, seemed to bridge this gap between fish and tetrapods, as they have 'internal nostrils' in their palate that superficially resemble the choanae. Early evolutionary anatomists thus considered that lungfishes constituted ideal intermediate forms, and were evidence for the piscine origin of tetrapods. Embryological studies then showed that these 'internal nostrils' in fact corresponded to the posterior nostrils of other fishes, but they were also taken to be homologous with — that is, the same as — the tetrapod choanae. The choanae were thus regarded as posterior nostrils that had migrated into the palate in the common ancestor of lungfishes and tetrapods.

Early in the twentieth century, however, that view was shaken by work on certain lobe-finned fishes from the Palaeozoic era (385–280 million years ago). These fishes, later referred to as 'osteolepiforms' and including such examples as *Osteolepis* and *Eusthenopteron*², proved to have a single external nostril on either side of the snout, and choana-like openings in the palate. They became preferred as 'ancestors' to tetrapods because of their close overall resemblance to them and their antiquity, and it soon became received wisdom that the tetrapods had originated from osteolepiforms. The lungfish 'internal nostrils' became increasingly regarded as having merely paralleled the tetrapod choanae in evolution, because of major differences in their relationships to the surrounding nerves, bones and sensory-line canals³. This was later confirmed by the discovery of *Diabolepis*, a primitive, 400-million-year-old lungfish, which retains two pairs of external nostrils (and thus lacks 'internal nostrils')⁴.

Several theories have been proposed for the origin of the choanae of tetrapods and of their fossil osteolepiform relatives. The two main possibilities are that choanae are derived from the posterior external nostrils, but independently of development of the 'internal nostrils' of lungfishes, or that they are a novelty, and the posterior external

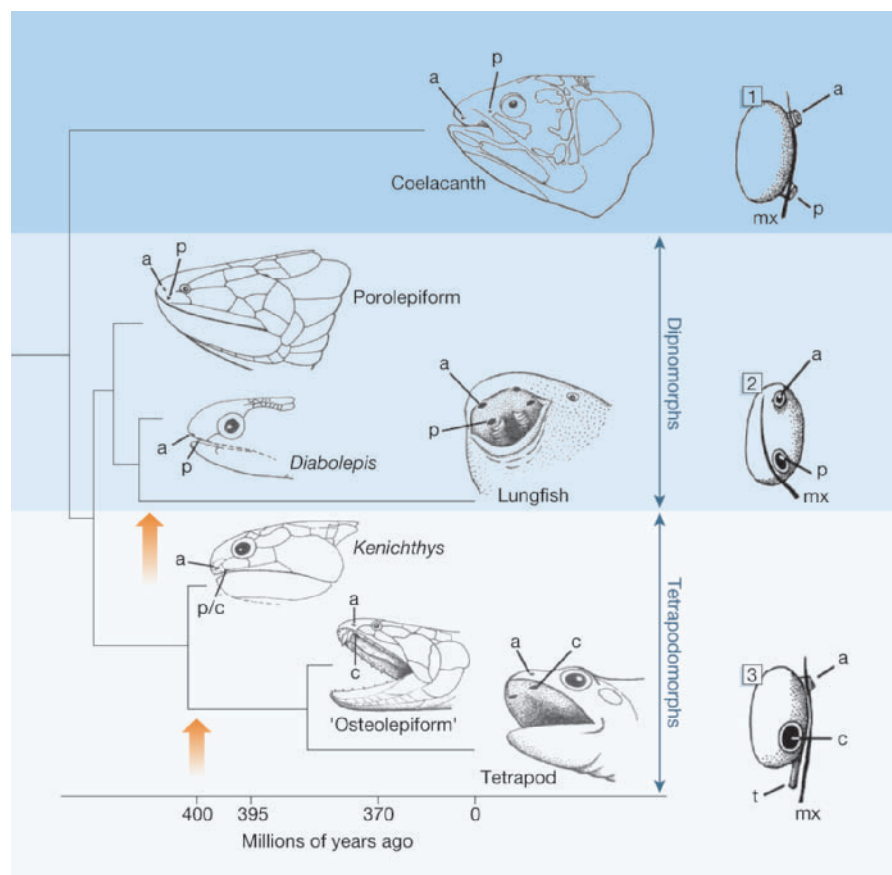


Figure 1 The hole picture. These depictions of the head and palate in some living and fossil lobe-finned fishes and tetrapods show the apertures of the nasal capsule — the anterior (a) and posterior (p) nostrils and the choanae (c). The tree indicates the relationships of the groups, and the orange arrows indicate the two occurrences of the migration of the posterior nostril into the palate. In *Kenichthys*, the fossil studied by Zhu and Ahlberg¹, the posterior nostril forms a notch in the outer dental arcade, and is intermediate in position between that of an external nostril and that of choanae (p/c). The diagrams (1–3) on the right are of the left nasal cavity, in ventral view, in the three living forms. In lungfishes, the path of the maxillary nerve (mx) is medial to the posterior nostril; in tetrapods it is lateral to the choana. t, tear duct.

nostrils have either disappeared or became the tear ducts. Debate on the matter rumbled on for decades. Rosen *et al.*⁵ even revived the theory that lungfishes have true choanae, and claimed that osteolepiforms provide no evidence for choanae, thereby creating a controversy⁶ that has had far-reaching consequences for the use of fossil data in reconstructions of evolutionary history⁷.

The current, morphology-based scheme of lobe-finned-fish and tetrapod evolution, first proposed by Ahlberg⁸ and shown in Figure 1, displays two major sister groups. These are the dipnomorphs (lungfishes, and the extinct *Diabolepis* and porolepiforms) and the tetrapodomorphs (tetrapods and their closest fossil fish relatives, including the osteolepiforms). If it is now clear, thanks to *Diabolepis*, that the migration of the posterior nostrils into the mouth occurred in the dipnomorphs within the lungfish lineage, there was no comparable fossil evidence for any precursor of the tetrapodomorph choanae.

That changes with Zhu and Ahlberg's report¹ on *Kenichthys*. *Kenichthys* is clearly a tetrapodomorph, and it displays exactly the intermediate condition that would be predicted when assuming that the tetrapod choanae were posterior nostrils displaced into the palate, in the same way as in lungfishes. The posterior nostrils (or already choanae?) of *Kenichthys* are actually right at the margin of the upper jaw, interrupting the outer dental 'arcade' formed by the premaxillary and maxillary bones.

This remarkable discovery, however, does not answer a question that was raised long ago³ and that centres on other anatomical features. If the posterior nostrils passed into

the palate to form the tetrapod choanae, why did they not leave indelible traces by interrupting the outer dental arcade and the sensory-line canal that runs below the eye sockets, and pushing the maxillary nerve inwards, as they did in lungfishes (Fig. 1)? Zhu and Ahlberg's analysis suggests that this canal and nerve, and the outer dental arcade, were restored laterally to the choanae in more advanced tetrapodomorphs. The restoration of the dental arcade may not be a problem, but that of the maxillary nerve and sensory-line canal certainly is. Further study of the nerve canals inside the snout of *Kenichthys*, or of tetrapod snout development, might offer a solution.

Nevertheless, *Kenichthys* clearly provides the first factual basis for the theory that tetrapod choanae actually are the posterior nostrils, as are the internal nostrils of lungfishes. In a way, the new analysis reconciles Rosen and colleagues' provocative theory⁵ and the classical interpretation of the fish members of the tetrapodomorphs — internal nostrils of lungfishes and tetrapod choanae are homologous, but their position in the palate is not. ■

Philippe Janvier is at the CNRS Unité Paléobiodiversité, Muséum National d'Histoire Naturelle, 8 Rue Buffon, Paris 75005, France, and the Natural History Museum, London. e-mail: janvier@mnhn.fr

1. Zhu, M. & Ahlberg, P. E. *Nature* **432**, 94–97 (2004).
2. Ahlberg, P. E. & Johanson, Z. *Nature* **395**, 792–794 (1998).
3. Jarvik, E. *Zool. Bidrag Uppsala* **21**, 235–675 (1942).
4. Chang, M. M. *Bull. Mus. Natl. Hist. Nat.* **17**, 235–268 (1995).
5. Rosen, D. E., Forey, P. L., Gardiner, B. G. & Patterson, C. *Bull. Am. Mus. Nat. Hist.* **167**, 157–276 (1981).
6. Jarvik, E. *Syst. Zool.* **30**, 378–384 (1981).
7. Gee, H. *In Search of Deep Time* (Free Press, New York, 1999).
8. Ahlberg, P. E. *Zool. J. Linn. Soc.* **103**, 241–287 (1991).

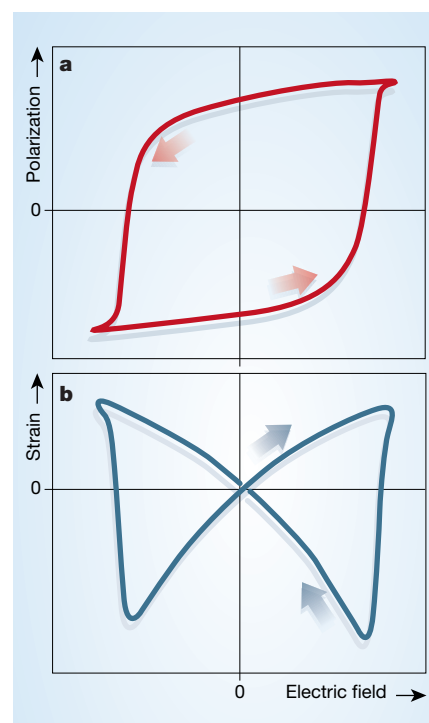


Figure 1 Hysteresis and shape change.

a, Hysteresis loop for a ferroelectric composition of lead zirconate and lead titanate. As the electric field is cycled, the polarization of the material changes, but differently depending on whether the field is increasing or decreasing. For zero electric field, there is remanent polarization; to force zero polarization a certain strength of coercive electric field is required. b, For a ferroelectric lead zirconate–titanate ceramic, a cycling electric field changes the elastic strain in the piezoelectric material. The slope of the curve at the origin is indicative of the high ' d_{33} ' of the ceramic — the induced charge per unit force applied in the same direction — and changes sign with the cycling of the electric field.

Materials science

Lead-free at last

Eric Cross

The most successful piezoelectric ceramics are based on lead zirconate and lead titanate. Environmental concerns over their lead content could disappear with the advent of a new ceramic that is lead-free.

Piezoelectricity is a time-honoured study in crystal physics, initiated by the brothers Jacques and Pierre Curie¹ in 1880. When a piezoelectric crystal is subjected to a suitably oriented elastic stress, the crystal polarizes electrically in proportion to the applied stress; conversely, when a suitably oriented electric field is applied, the crystal changes shape (strains) in proportion to the field level. Such properties lead naturally to possibilities for sensing and actuating elastic changes, but in simple piezoelectrics the effects are too small. Compositions of lead zirconate titanate — the 'PZT' family of ceramics — show much stronger piezoelectric effects, and have come

to dominate the field. Their lead content, however, raises environmental concerns. So the advance reported by Saito *et al.*² on page 84 of this issue is welcome indeed: they have created lead-free piezoceramics with properties that closely match those of PZT.

In a normal ceramic, the random orientation of the individual crystallites imparts an infinite degree of rotational symmetry within the ceramic texture; as a result, irrespective of the crystallite symmetry, piezoelectricity is forbidden. What makes piezoelectricity possible in some ceramics, and markedly raises the level of piezoactivity in some single crystals, is the phenomenon of ferroelectricity. In the ferroelectric crystal or crystallite,

there is a substructure of electrically polar domains that can be reoriented by a strong applied electric field. This domain reorientation is demonstrated by the appearance of electric hysteresis and significant shape change (Fig. 1).

In spite of the fact that hysteresis in ceramic ferroelectrics was well known to the physics community, it was an engineer in the United States, R. B. Gray, working at Erie Technological Products in Erie, Pennsylvania, who realized that ferroelectric capacitors under test there were humming at the common 60-Hz power-line frequency. From this he deduced their piezoelectricity, made for himself a piezoceramic phonograph pick-up and pinned down the master piezoceramic patent³ for Erie. The patent, subsequently licensed to the Clevite Corporation, established an early lead for the United States in the piezoceramic business, and this was strongly reinforced by Clevite's role in the development of the now-dominant PZT family of piezoceramics.

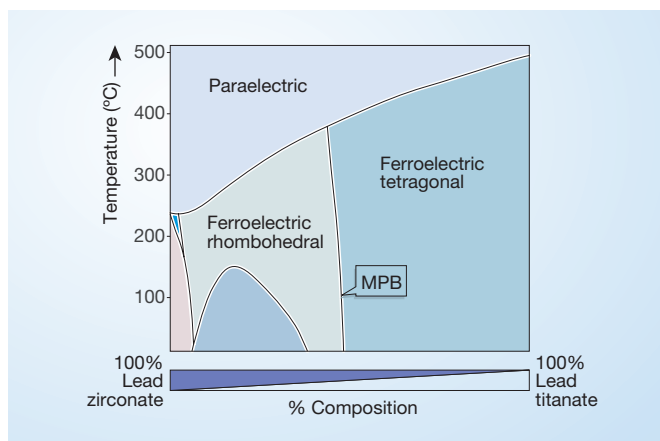


Figure 2 Phase diagram for the lead zirconate-titanate ceramic. Different regions correspond to different crystallite structures. Around the morphotropic phase boundary (MPB), at a composition of 52% lead zirconate to 48% lead titanate, the piezoelectric properties of the ceramic are particularly attractive.

PZT compositions had first been explored in Japan^{4,5}. But it was Jaffe *et al.*⁶, at the US Bureau of Standards (now the National Institute of Standards and Technology), who uncovered the vertical 'morphotropic phase boundary' and realized its likely role in enhanced piezoelectric activity. Depending on the proportions of lead zirconate and lead titanate it contains, and the temperature, a PZT composition has a particular crystal structure, as indicated in the phase diagram of Figure 2. The morphotropic phase boundary, positioned where the zirconate-to-titanate ratio is 52:48, marks the transition from rhombohedral to tetragonal ferroelectric crystallites.

Because the polar vector of domains changes orientation spontaneously when the ferroelectric phase boundary is crossed, for compositions close to the boundary it is quite easy for an electric field to tilt the polar vector. This gives rise to very strong piezoelectric activity in the ceramic. That the phase boundary at the 52:48 Zr/Ti composition in PZT is nearly vertical (morphotropic) in the phase diagram means that these enhanced values persist over a usefully wide temperature range.

Clevite was quick to exploit the substantial advantages of the PZT system over the other ferroelectric ceramics then available⁷. These included the high activity of the lead cation, which made it possible to process the material reproducibly using conventional methods.

Since then, both suppliers and markets for piezoceramics have diversified markedly, with the centre of gravity moving steadily towards Asia. High levels of piezoelectric activity had, however, remained confined to lead-based systems, with PZT still the major 'workhorse'. Piezoelectric sensing and actuation have become a reality in many systems. The US Navy is a major customer, for sonar, sonobuoys and the whole-ocean security network. PZT ceramic-polymer composites, originally developed for navy systems and now miniaturized, are essential elements in medical ultrasound tomography. There are innumerable applications in cars, including new multi-pulse quieting fuel injectors

for small diesel engines. More recently, the area of 'smart' materials has blossomed, imparting the capabilities of solid-state electronics to structural materials through the incorporation of sensors, actuators and suitable electronic feedback and control.

The only drawback in the whole technology was the environmentalist's nightmare of its dependence on a 60%-lead-containing family of ceramics. Now at last there is light at the end of that tunnel. Saito and colleagues² in Japan have found an alternative, in the shape of texturally modified sodium potassium niobate ceramic composition. For this material, there is a morphotropic phase boundary between tetragonal and orthorhombic ferroelectric forms (Fig. 1a on page 85). For a composition close to this boundary, the material's properties are comparable to those of a basic, unmodified PZT ceramic, and its Curie temperature — at which it loses its ferroelectricity — is equivalently high. What is even more striking is that, by a tractable process, Saito *et al.* have gone further to produce excellent grain-oriented ceramics, with enhanced properties that match those of an optimum modified PZT composition. For one measure in particular, the new lead-free ceramic even surpasses PZT: the induced charge per unit force (applied in the same direction; a quantity known as d_{33}) is more stable against changes of temperature in the lead-free variety.

This demonstration should certainly refocus efforts on the topic of lead-free piezoceramics. As the results are brought to the market place, all producers may be forced to explore this option.

Eric Cross is in the Materials Research Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802, USA.
e-mail: lec3@psu.edu

1. Curie, J. & Curie, P. *Bull. Soc. Fr. Mineral.* 3, 90 (1880).
2. Saito, Y. *et al. Nature* 432, 84–87 (2004).
3. Gray, R. B. US Patent No. 2,486,560 (1949).
4. Shirane, G. & Takeda, A. *J. Phys. Soc. Jpn* 7, 5 (1952).
5. Sawaguchi, E. *J. Phys. Soc. Jpn* 8, 615–629 (1953).
6. Jaffe, B., Roth, R. S. & Marzullo, S. *J. Appl. Phys.* 25, 809–810 (1954).
7. Jaffe, B., Cook, W. R. Jr & Jaffe, H. *Piezoelectric Ceramics* (Academic, London, 1971).



100 YEARS AGO

Mr. H. G. Wells returns to the more serious side of his work in "A Modern Utopia," which is being published month by month in the *Fortnightly Review*. As in "Anticipations" and "Mankind in the Making," Mr. Wells concerns himself with sociological problems, and pictures the probable manners and customs of society in a Utopia, situated on a distant planet, which is the natural outcome of continued development on modern lines.

The *Journal of Anatomy and Physiology* for October (xxxix., part i.) contains a number of valuable papers, but of purely anatomical interest. The principal contribution is by Dr. Huntington on the derivation and significance of certain supernumerary muscles of the pectoral region, illustrated with fourteen excellent coloured plates.

Messrs. Routledge and Sons, Ltd., have added to their series of "Country Books" a profusely illustrated edition of Charles Kingsley's "Glaucus, or the Wonders of the Seashore." The volume is published at 3s. 6d. From *Nature* 3 November 1904.

50 YEARS AGO

Although it is impossible for one group of workers to examine more than a few fundamental features of so complex a biological system, I wish to emphasize the need for an integrated outlook if we are to reach a proper understanding of the nature, structure and function of connective tissues. The physicist, in thinking of connective tissues, almost automatically concentrates his attention on the structural properties of the fibrous collagen; the biochemist likewise tries to characterize the protein by a precise analysis or to determine, perhaps, what sugars are present in the ground substance. The histologists and pathologists are aware more than any of us that these are but facets of the whole problem. Collagen fibres are a characteristic feature of all connective tissues and their derivatives... But all connective tissues are composed of four main constituents: cells, fibres, matrix or ground substance, and intercellular fluid. It is too often forgotten that the cell is the focus and origin of all tissues. If we are to reach a better understanding of the collagen system, more attention must be paid to the fibrogenic cells, to which the fibres and ground substance in all probability owe their origin.

J. T. Randall

From *Nature* 6 November 1954.

Cell biology

Adhesion articulated

A. Paul Mould and Martin J. Humphries

A new structure of the 'head' region of an integrin protein explains the remarkable vertical extension that enables these molecules to rise to the task of mediating cell adhesion.

If they are to survive and move, cells must be able to stick to their surroundings. Integrin proteins enable a bidirectional control of cell adhesiveness, by dynamically coupling the matrix of molecules found outside the cell to the cell's internal 'skeleton'. All 24 members of the integrin family have two subunits, one from the α -subfamily and one from the β -subfamily; each subunit is composed of large extracellular and short intracellular regions. Integrins work by conveying shape changes through these subunits — from the intracellular domains, through the cell membrane, to the extracellular region, and vice versa^{1,2}. As adhesion needs to be strictly controlled to avoid inappropriate cell aggregation or migration, integrins require a highly responsive structure that can exist in low-affinity (inactive), high-affinity (primed) or ligand-bound (activated) states.

On page 59 of this issue, Xiao *et al.*³ explain the structural basis of intramolecular signal transduction for one particular integrin, termed $\alpha_{IIb}\beta_3$, which is found on platelets and binds to the blood protein fibrinogen. In doing so, the authors reveal the core mechanism that controls cellular adhesiveness.

The first integrin crystal structure (of the extracellular domains of integrin $\alpha_v\beta_3$, a receptor for the extracellular-matrix protein vitronectin) revealed an anthropomorphic molecule composed of a ligand-binding

'head' and two long 'legs'⁴. Each subunit (α and β) contributed one of the legs, and both subunits combined to form the head. Interestingly, the legs were highly bent at their knees, such that the head was very close to the lower legs (Fig. 1a). In this structure, some of the modules making up the knee regions were unresolved (including the so-called PSI domain), suggesting that these regions were highly flexible.

Although a huge advance for the field, these findings presented a conundrum: how can shape changes be transduced from an integrin's head to its legs, or from its legs to its head, particularly given that the knees are wobbly-looking? A controversy immediately arose as to whether the bent structure was physiologically relevant, and if so whether it represented a low-affinity or a high-affinity state⁵. Subsequently, several studies showed that the bent form can undergo a dramatic straightening to an extended form (Fig. 1b)^{6–8}.

Now Xiao *et al.*³ have crystallized a fragment of the head region of integrin $\alpha_{IIb}\beta_3$ in complex with an antibody fragment and with one of three different compounds that mimic its natural ligand (fibrinogen), two of which are used in the clinic for treating clotting in the coronary artery. These peptide-bound structures (which look very similar) represent the ligand-activated state (Fig. 1c). Although these structures lack most of the leg regions, major alterations in the structure

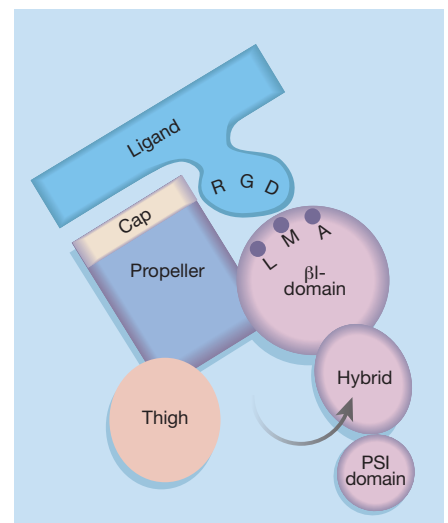


Figure 2 The complex between a macromolecular ligand and the integrin head domains. The cap, propeller, β I, hybrid and PSI domains are present in the structure described by Xiao *et al.*³. The position of the thigh can be inferred from a previous structure⁴. Ligand binds across both subunits (α on the left; β on the right), with the RGD (arginine–glycine–aspartate) sequence within the ligand binding at the junction between the α - and β -subunits. This sequence also coordinates the cation present in the site known as MIDAS (metal-ion-dependent adhesion site; M) of the integrin. Secondary sites in the ligand contact the α -subunit cap region. Sites in the integrin that are known as ADMIDAS (adjacent to MIDAS, A) and LIMBS (ligand-induced metal-binding site, L) also bind cations and control the protein's activation state. Ligand binding stabilizes an outward swing (arrow) of the rigidly connected hybrid and PSI domains, and thereby separates the integrin legs (not shown).

of the head are apparent compared with the $\alpha_v\beta_3$ structure, and the likely effect of these changes on the overall shape of the intact molecule can be inferred.

So, what do we learn? The most obvious difference between the inactive $\alpha_v\beta_3$ and active $\alpha_{IIb}\beta_3$ structures is a large outward swing of a module in the β -subunit known as the hybrid domain. This domain, which is located at the 'neck' of the integrin, undergoes a reorientation of some 60° relative to the ligand-binding β I-domain. The β -subunit's PSI domain is observed to lie below the hybrid domain and to act as a rigid connecting rod that translates the swing of the hybrid domain into a separation of the leg regions. Thus, the integrin knees are not wobbly but are instead precise articulation points.

These findings strongly suggest that the inactive form of an integrin is bent and the fully active form is unbent, with an altered arrangement of the head domains. Such a model fits well with the extended forms of integrins observed by electron microscopy⁷, and with the exposure of particular sites (detected by various antibodies) on the

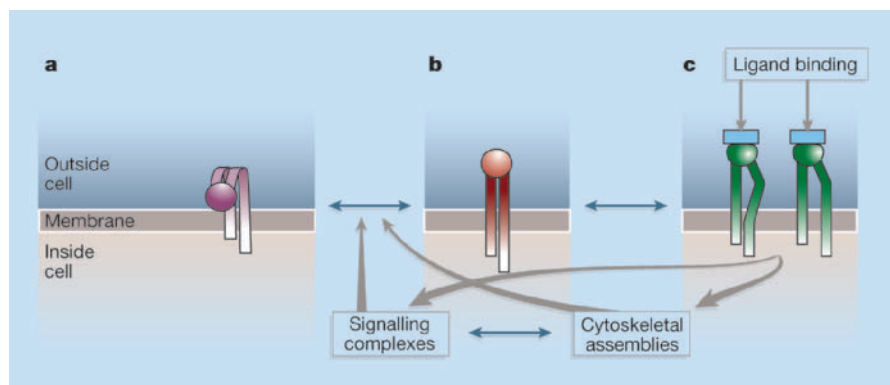


Figure 1 Integrin activation states. This schematic representation shows bent (a), extended (b) and ligand-bound (c) forms of an integrin protein, inserted into the plasma membrane of a cell. The conformational equilibria between each state are regulated either by the binding of cytoskeletal or signalling factors to the intracellular domains (below the membrane) or by ligand binding (blue rectangles). a, In the bent form, the head region points inwards towards the cell surface, and has low affinity for other molecules. b, In the extended form, the head region points away from the surface, allowing it to engage other molecules; the different domains that make up the head region also alter its shape (not shown), increasing the affinity of ligand binding. c, Ligand engagement stabilizes an outward swing of the β -subunit (right-hand leg) and potentially allows variable leg separation.

hybrid domain and knee regions that occurs when integrins are in a high-affinity state^{6,9}. Movement of the hybrid domain is also linked to changes in the shape of the neighbouring β I-domain. These changes allow the head region to undergo a metaphorical facelift so that ligands can bind with higher affinity. The major recognition site for ligands, including the classical arginine-glycine-aspartate amino-acid sequence found in many extracellular proteins, is in a groove between the α - and β -subunits (as shown¹⁰ for integrin $\alpha_v\beta_3$); however, the new structure reveals a larger subregion of the α -subunit that provides a secondary ligand-binding site (Fig. 2).

Although many of the long-standing mysteries of integrin structure have now been solved, several gaps remain. For instance, there is still a need to obtain a co-crystal between an integrin and a larger (macromolecular) ligand fragment, so that the ligand-binding pocket can be completely defined. Furthermore, nine integrins contain an additional ligand-binding region, called the α A- or I-domain, inserted into the head. Solving the structure of one of these integrins would explain the atomic basis of communication between this domain and the β I-domain.

In addition, the paper by Xiao *et al.*³ raises several specific questions. First, given that different $\alpha_{\text{IIB}}\beta_3$ -binding therapeutic peptides stabilized similar integrin conformations, how is the binding of different ligands to integrins converted into a graded signal inside the cell? One possibility is that variable 'outside-in' signalling is determined by differences in the kinetics of leg separation in different ligand-integrin complexes. Alternatively, there might be situations in which different ligands stabilize conformations with subtle differences in the angles of the joints (Fig. 1c). Determining the atomic basis of knee flexing will help in this regard. Second, to what extent are the same conformational changes involved in inside-out and outside-in signal transduction? In particular, to what extent does inside-out signalling cause unbending or the hybrid domain to swing out?

Finally, inappropriate integrin-ligand interactions contribute to aberrant cell adhesion in many diseases, including inflammation, the blockage of blood vessels by blood clots, and tumour progression. Can the availability of structures of integrin $\alpha_{\text{IIB}}\beta_3$ in complex with therapeutic peptides aid the rational design of drugs to target other integrins? Perhaps: it is certainly likely that the strategy of crystallizing a truncated form of $\alpha_{\text{IIB}}\beta_3$ could also be used for other integrins (including $\alpha_4\beta_1$ and $\alpha_v\beta_3$) that are major therapeutic targets in rheumatoid arthritis, multiple sclerosis and cancer. ■

A. Paul Mould and Martin J. Humphries are at the Wellcome Trust Centre for Cell-Matrix Research, Faculty of Life Sciences, University of Manchester,

Manchester M13 9PT, UK.

e-mail: martin.humphries@man.ac.uk

1. Carman, C. V. & Springer, T. A. *Curr. Opin. Cell Biol.* **15**, 547–556 (2003).
2. Mould, A. P. & Humphries, M. J. *Curr. Opin. Cell Biol.* **16**, 544–551 (2004).
3. Xiao, T., Takagi, J., Collier, B. S., Wang, J.-H. & Springer, T. A. *Nature* **432**, 59–67 (2004).

4. Xiong, J.-P. *et al. Science* **294**, 339–345 (2001).
5. Liddington, R. C. *Structure* **10**, 605–607 (2002).
6. Beglova, N. *et al. Nature Struct. Biol.* **9**, 282–287 (2002).
7. Takagi, J., Petre, B. M., Walz, T. & Springer, T. A. *Cell* **110**, 599–611 (2002).
8. Chigayev, A. *et al. Biophys. J.* **85**, 3951–3962 (2003).
9. Mould, A. P. *et al. J. Biol. Chem.* **278**, 17028–17035 (2003).
10. Xiong, J.-P. *et al. Science* **296**, 151–155 (2002).

Land management

Forests, fires and climate

Cathy Whitlock

A new analysis of the effect of climatic variation on forest fires goes back several thousand years. One take-home message is that a one-size-fits-all forest management strategy is, literally, short-sighted.

Seemingly unprecedented events in human lifetimes can be business-as-usual when viewed on longer timeframes. But that's not always recognized. For example, management strategies in the United States that seek to restore landscapes to the conditions that prevailed at the time of first European contact often fail to consider the events that created those conditions. This short-sightedness is particularly evident in the area of wildfire management.

On page 87 of this issue, however, Pierce *et al.*¹ provide a long-term perspective on the effects of fire in 'low-elevation' pine forests in the western United States. These dry, open forests lie above the steppe and grasslands that occupy most intermontane basins in the Rocky Mountains; in more moist settings at higher elevations, they are replaced by closed montane and subalpine forests. The low-elevation forests have been extensively modified by human activities, and there is intense debate about the appropriate action needed to restore them to their prehistoric state².

In the past 15 years, the western United States has experienced some extreme fires, notable for their size and severity. The annual costs of fire suppression now exceed \$1.6 billion, and the ceiling seems nowhere in sight³. In the absence of large fires during most of the twentieth century, many forests have become filled with a dense understory of shrubs and small trees that provide 'ladder fuels' that set the crowns of trees alight: these crown fires are the most destructive types of wildfire. The Healthy Forest Restoration Act⁴, signed into law by President George W. Bush in 2003, purportedly seeks to redress the ecological effects of fire suppression by establishing programmes of aggressive thinning, deliberate burning, and replanting to create open conditions. For local communities with economies based on timber extraction, this law is good news; for environmentalists, it is a travesty that limits scientific analysis and public participation in decision-making and policy.

But are the fires of the past 15 years



Figure 1 Slide show. This slope failure occurred following rainfall on melting snow in 1997, in an area of the South Fork Payette River, Idaho, that was severely burned in 1989. Herbs and low shrubs had regrown after the fire. But the decay of tree roots probably caused the fatal weakening of the slope.

G. A. MEYER

unprecedented in their severity? Another consideration in this debate is the unusual climate over that time that has dried fuels and created conditions conducive to the ignition and spread of fire⁵. Current fires, driven by strong winds, have burned both open and closed forests, ignoring conventional wisdom about fire behaviour. This conventional view is that closed, fuel-laden forests should experience severe stand-replacing fires, whereas open forests should be relatively fire resistant.

Understanding how present-day fires are influenced by climate requires information both on past fire conditions and on climate variability. If fire activity is a predictable consequence of the interconnectedness of climate, fire and vegetation, then in the Rocky Mountains we can expect fire regimes typical of a warmer, drier climate in the future⁶.

Our knowledge of fire history in low-elevation pine forests is based largely on analyses of fire scars on living and dead trees⁷. Most tree-based studies offer information about the past 500 years or so, but are biased towards surface fires that scar, but do not kill, trees. Charcoal records from lake sediments extend the fire chronology back several millennia; however, these studies generally focus on montane and subalpine forests⁸.

Pierce *et al.*¹ offer an alternative data set for studying past fire regimes at low elevations — the record of fire-related sedimentation, preserved as accumulations of rock, soil and wood known as debris-flow deposits. Severe fires can remove ground litter and reduce infiltration of water; following that, storms and melting snow in spring can initiate landslides that deposit fire-related sediment in valley bottoms⁹ (Fig. 1). In contrast, the geomorphological effects of low-severity surface fires are less dramatic, because of a more discontinuous burn pattern and lesser impact on the soil.

Pierce *et al.* determined the age of charcoal particles in different types of debris-flow deposits, and established a chronology of fire-related erosion for the past 8,000 years. Their data suggest that fire-related debris flows were associated with warm, dry periods, such as the Medieval Climate Anomaly (roughly AD 950–1350), when grass cover was sparse and fires were severe. In contrast, during cooler, wetter periods, such as the Little Ice Age (AD 1350–1900), dense grass cover supported frequent surface fires and there were only comparatively minor episodes of sedimentation.

Two points that emerge from this study¹ are worth mentioning. First, fire is influenced by climate variations that occur on several timescales, and shifts in fire regimes affect the various factors — vegetation, and hydrological and geomorphological processes — that stabilize or destabilize landscapes. In a given year, fire occurrence and severity are governed by particular

weather patterns and their influence on fuel moisture, ignition conditions and fire behaviour¹⁰. Fire-related geomorphological events are determined by the fire intensity, as well as the subsequent weather and the characteristics of the mountain slopes involved. On inter-annual timescales, fuel availability and slope stability are affected by variations in climate arising from atmosphere–ocean interactions, such as the El Niño–Southern Oscillation. On longer timescales, alternation of wet and dry periods during past centuries and millennia has shifted pine-forest ecosystems between surface-fire and crown-fire regimes and, in turn, has shifted the character of post-fire sedimentation.

Second, recent fires in low-elevation forests near sizeable human populations have led to calls for draconian tree and understorey thinning. Yet the investigations of Pierce *et al.*¹ and tree-ring studies^{11–13}, suggest that fire activity in these forests has varied in the past and includes episodes of severe crown fires and large debris flows. We should consider this long-term perspective before embracing one-size-fits-all management strategies. And in the future, we need to be guided by a firm understanding of both the relationship

between fires, climate and landscape processes on different timescales, and the interactions between climate and land-management activities in shaping fire regimes.

Cathy Whitlock is in the Department of Earth Sciences, Montana State University, Bozeman, Montana 59717, USA.

e-mail: whitlock@montana.edu

1. Pierce, J. L., Meyer, G. A. & Jull, A. J. T. *Nature* **432**, 87–90 (2004).
2. DellaSala, D. A., Williams, J. E., Williams, C. D. & Franklin, J. F. *Conserv. Biol.* **18**, 980–986 (2004).
3. National Interagency Fire Center www.nifc.gov (2004).
4. Healthy Forests Restoration Act of 2003 www.healthyforests.gov/initiative/legislation.html (2003).
5. Westerling, A. L., Gershunov, A., Brown, T. J., Cayan, D. R. & Dettinger, M. D. *Bull. Am. Meteorol. Soc.* **84**, 595–604 (2003).
6. McKenzie, D., Gedalof, Z., Peterson, D. L. & Mote, P. *Conserv. Biol.* **18**, 890–902 (2004).
7. Agee, J. K. *Fire Ecology of Pacific Northwest Forests* (Island Press, Washington DC, 1993).
8. Whitlock, C. & Bartlein, P. J. in *The Quaternary Period in the United States* (eds Gillespie, A. R., Porter, S. C. & Atwater, B. F.) 479–490 (Elsevier, Amsterdam, 2004).
9. Cannon, S. H., Bigio, E. R. & Mine, E. *Hydrol. Processes* **15**, 3011–3023 (2003).
10. Pyne, S. J., Andrews, P. L. & Laven, R. D. *Introduction to Wildland Fire* (Wiley, New York, 1996).
11. Grissano-Mayer, H. D. & Swetnam, T. W. *Holocene* **10**, 213–220 (2000).
12. Veblen, T. T., Kitzberger, T. & Donnegan, J. *Ecol. Appl.* **10**, 1178–1195 (2000).
13. Brown, P. M., Kaufmann, M. R. & Shepperd, W. D. *Landscape Ecol.* **14**, 513–532 (1999).

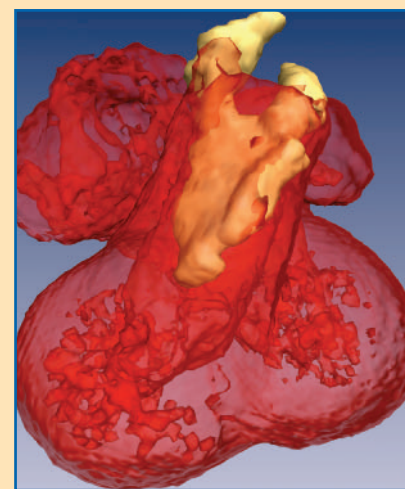
Cardiovascular biology

How genes know their place

Embryonic development is largely a matter of switching on the right genes, in the right place, and at the right time — it's no use activating heart-manufacturing genes in the limbs, for instance. Elsewhere in this issue (*Nature* **432**, 107–112; 2004), Benoit G. Bruneau and colleagues describe a protein that keeps heart-specific genes in their place.

Using mice, the authors first discovered that the protein in question, named Baf60c, is initially expressed only in the developing heart. Investigating further, they found that completely eliminating Baf60c from mouse embryos led to major cardiac defects (and early death). Knocking out about 50% of the protein led to somewhat milder, although still ultimately fatal, problems, such as an abnormally straight and split outflow tract (the yellow area in this picture).

Why do these problems occur? Baf60c is part of a complex that controls the accessibility of genes by interacting with transcription factors — proteins that regulate gene expression. So perhaps, without Baf60c, heart-specific genes cannot gain access to the transcriptional machinery and therefore remain inactive. Indeed, Bruneau and colleagues found that many genes that are characteristic of different aspects of heart development are not expressed when Baf60c is eliminated, including various genes involved in the formation of the outflow tract. Moreover, in *in vitro* studies, the authors showed that Baf60c enhances the interaction between Brg1, a key component of the complex in which Baf60c is found,



and certain heart-specific transcription factors.

So, at early stages of development, Baf60c seems to ensure that a complex that alters the structure of DNA — and hence its accessibility to proteins that activate genes — is targeted specifically to genes that are needed for heart development. This mechanism might also work in other tissues, with proteins other than Baf60c providing the necessary specificity. And the findings might bear on human disease: the defects caused by a partial lack of Baf60c are somewhat similar to those seen in some people with congenital heart defects. It remains to be seen, however, how the heart-specific expression of Baf60c is achieved in the first place.

Amanda Tromans

J. R. WALLS

Obituary

Fred Lawrence Whipple (1906–2004)

Fred Whipple died on 30 August, two months shy of his ninety-eighth birthday. The breadth of his published research, from 1927 to 2000, is extraordinary. It covers such diverse topics as comets, meteors, satellite tracking, variable stars, supernovae, stellar evolution, radio astronomy and astronomical instrumentation. Whipple's collected works were published in two massive volumes in 1972, shortly before his 'retirement'. But his research contributions continued for another three decades: an additional volume is planned.

Fred Lawrence Whipple was born on 5 November 1906, on a farm in Red Oak, Iowa. When he was 15, the family moved to California. There, Whipple studied mathematics at Occidental College and at the University of California at Los Angeles. As a graduate student at the University of California at Berkeley in 1930, he was one of the first to compute the orbit of the newly discovered planet Pluto. On receiving his PhD in 1931, he joined the staff of the Harvard College Observatory in Cambridge, Massachusetts. He became chairman of Harvard University's astronomy department in 1949, then director of the Smithsonian Astrophysical Observatory from 1955 until 1973.

At Harvard, Whipple developed a photographic tracking network of professional and amateur observers with the aim of determining the trajectories of meteors from two or more simultaneous observations. To fill the daytime gap when meteors could not be photographed, Whipple organized another programme for the detection of these objects using radio-wavelength observations. Eventually, in collaboration with Richard McCrosky and others, he concluded that most meteors were on comet-like orbits; fewer than 1% of the sporadic meteors visible to the naked eye could be traced to an origin outside the Solar System. During his career, he discovered six comets. He liked to quip that anyone could discover a comet — all it took was time. He also discovered the asteroid 1252 Celestia, which he named after his mother; asteroid 1940 was renamed 1940 Whipple, in his honour.

When the Soviet satellite Sputnik was launched in 1957, Whipple's band of meteor observers was the only network in place that could track its progress visually. Later, he developed a photographic tracking system for meteors and artificial satellites that proved so successful and precise that the satellite-tracking data



A founding father of modern planetary science

could be used to model the variation in Earth's shape and density from the observed gravitational effects on the orbits. He once noted that the highlight of his career was having his family and parents present at the White House when in 1963 he received the President's Award for Distinguished Public Service for this work, from John F. Kennedy.

His seminal work in cometary science, starting in 1950, was the 'dirty snowball' model for the nuclei of comets. It prompted a paradigm shift. A comet had been thought to be a flying cloud of particles. Whipple instead envisaged the cometary nucleus as a conglomerate of ices (mostly water, ammonia, methane, carbon dioxide and carbon monoxide) embedded in a non-volatile matrix of meteoric material. Part of his rationale was to provide an explanation of the so-called non-gravitational forces acting on comets: the rocket-like thrusting of a comet when the ices vaporize near the Sun introduces a small but noticeable thrust on the comet itself. With this effect properly modelled, the motions of active comets could be predicted far more accurately.

Later spacecraft observations showed that comets are surrounded by enormous atmospheres of hydrogen, confirming that the major cometary ice was likely to be water. In 1986, images from the Giotto spacecraft revealed that the nucleus of comet Halley was indeed a solid body made of dirty ice — a dramatic

confirmation of Whipple's model. He was fond of telling the story that, while working on his 'dirty snowball' papers, he couldn't remember the value for the tensile strength of water ice; he solved the dilemma in characteristic Whipple fashion, by picturing the longest icicle he had seen while growing up in Iowa.

During the Second World War, Whipple led an effort to develop a means of confusing enemy radar — strips of reflective aluminium (known as 'chaff') dropped from Allied aircraft was his solution, for which he received a certificate of merit from US President Harry S. Truman. Eleven years before the launch of Sputnik, he developed the Whipple Shield: a series of thin metallic layers that stands out from a spacecraft and protects it from high-speed interplanetary dust particles. When particles hit this outer layer, they fragment and vaporize, and the debris lacks the energy to penetrate the main spacecraft. This same design successfully protected the Stardust spacecraft during a comet flyby in January of this year.

Late in his career Whipple was responsible for the construction of a major optical observatory on Mount Hopkins in Arizona. In 1981, it was renamed the Fred Lawrence Whipple Observatory. Whipple was successful as both a manager of large science enterprises and as a researcher. One of his secrets, he said, which enabled him to do management and science simultaneously was to spend some mornings in a room adjacent to his office, doing research. His secretary was asked to (correctly) notify callers that "Dr Whipple is not in his office at the moment" and could he or she call again later in the day.

An innovative thinker and a major force in the growth of planetary science in the twentieth century, Whipple was always a gentleman and a friend to all his colleagues. He was just plain 'Fred' to all who knew him. He was keenly interested in what the younger generation of scientists was doing, and throughout his life he remained a conscientious mentor to his former students. When asked the secret of his longevity at his ninetieth birthday party, he noted, "you've to start early". Fortunately for planetary science, Fred Whipple did start early — and he stayed late.

Donald K. Yeomans and Joseph Veverka

Donald K. Yeomans is at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA.

e-mail: donald.k.yeomans@jpl.nasa.gov

Joseph Veverka is at the Center for Radio Physics and Space Research, and the Laboratory for Planetary Studies, Cornell University, Ithaca, New York 14850, USA.

e-mail: jfv4@cornell.edu

Medicine

Gene therapy for prion diseases?

J. Cell Sci. **117**, 5591–5597 (2004)

Transmissible spongiform encephalopathies — such as scrapie in sheep, mad cow disease and human Creutzfeldt–Jakob disease — are thought to be triggered by an aberrant, infectious form of prion protein. This converts other, healthy prion proteins into the infectious form, which accumulate and eventually kill neurons.

Carole Crozet *et al.* have found that a type of gene therapy may stop the aberrant prion protein from proliferating in infected cells. They based their technique on two naturally occurring, disease-resistant forms of prion protein. These proteins do not switch into the infectious form and are thought to protect some sheep and humans from brain-wasting disease.

The authors engineered mouse prion genes to be similarly disease-resistant, placed them in a lentiviral ‘carrier’, and allowed the virus to inject the genes into cultured mouse cells already infected with disease-causing prions. The cells started to manufacture the disease-resistant prion, and this stopped the infectious prion from multiplying.

Crozet *et al.* went on to show that the lentivirus, unlike many other viruses used in gene therapy, can infect brain cells that have ceased dividing. Gene therapy for prion diseases — for which there is currently no treatment — may one day prove a reality.

Helen Pearson

Molecular physics

Liquid mix

Science **306**, 848–851, 845–848 (2004)

There are many ways to stack atoms and molecules in a solid — ice has at least 12 crystalline forms. But a liquid is a liquid, a mere disordered jumble of particles. Or so one might think. Yet over the past few years there have been signs that some substances may have more than one liquid state. This has been claimed both for water and for a melt of Al_2O_3 – Y_2O_3 when they are in supercooled (and thus metastable) states.

Yoshinori Katayama *et al.* have previously reported signs of a first-order phase transition (like the jump from liquid to gas) between two stable liquid phases of phosphorus. Now they clinch the case by directly observing the coexistence of high- and low-density phases of phosphorus at 1,000 °C and high pressure. X-ray diffraction measurements show that the liquids have different structures.

Meanwhile, also in the latest issue of *Science*, Rei Kurita and Hajime Tanaka

add to the story. They report a transition between two liquid phases of the compound triphenyl phosphite.

Philip Ball

Development

Eye for stem cells

Proc. Natl Acad. Sci. USA **101**, 15772–15777 (2004)

The human retina maintains a small population of stem cells, say Brenda L. K. Coles *et al.*, and does so throughout life.

Coles and colleagues studied retinas taken from human cadavers ranging from newborns to subjects in their seventies. They isolated and cultured cells from various parts of the retina, and found that those from a region called the retinal ciliary margin had the hallmarks of stem cells: the ability to proliferate and self-renew, and (crucially) to develop into a range of different types. Each eye contained around 10,000 such cells.

When the researchers injected the stem cells into developing mouse or chick eyes, the cells’ progeny migrated to the retina and differentiated — most showing the characteristic markers of photoreceptor cells, the retina’s light-sensing apparatus. The authors argue that this shows that there is no qualitative difference between the retinal stem cells of different animal species. What’s more, the lifelong presence of these cells might mean they could be used in tackling the human retinal degeneration that is associated with ageing.

Michael Hopkin

Food science

Down the hatch

J. Agric. Food Chem. **52**, 6564–6571 (2004)

The difficulty of simulating our acutely tuned olfactory system isn’t the only reason that taste-testing of food and wine is hard to automate. Our sensory experience of foods depends not just on the complex blend of aromas that the food gives off, but on exactly how these are released in the mouth and throat. The heady effects of a fine whisky or a rich Cambozola seem to derive mostly from a thin film of liquid or semi-liquid material that is deposited in the throat during swallowing and which discharges its aroma molecules into the breath.

Koen G. C. Weel *et al.* have devised an artificial throat that can mimic this release process. The ‘throat’ is a simple piece of rubber tubing that bridges glass tubes above and below it. The upper tube can be filled with liquids, including saliva, while the rubber tube is clamped shut. To ‘swallow’, the clamp is opened and the liquids drain through, leaving a film on the rubber tube. Air is then pumped upwards to carry the aroma molecules to a mass spectrometer, which acts as the nose. The release patterns are generally similar to those measured in human tasters — except that this device could be used for unpalatable or even toxic substances that no taster could be expected to stomach.

Philip Ball

Glaciology

Under the ice

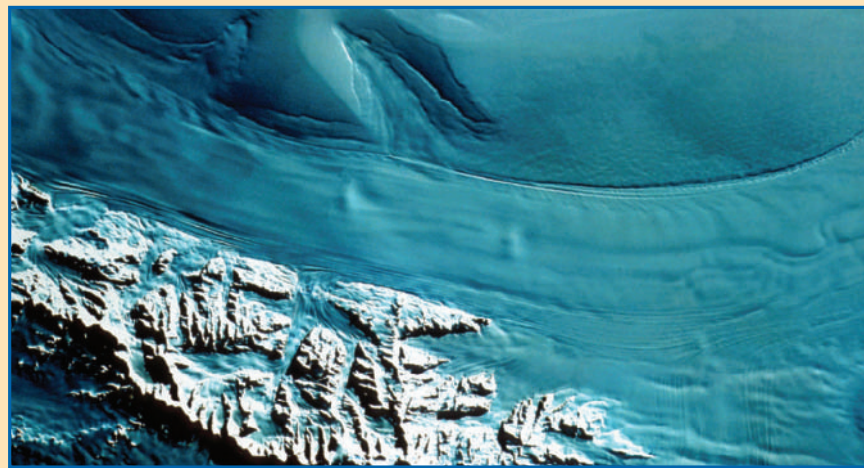
Geophys. Res. Lett. **31**, L20401 (2004)

Ice slips off the edge of Antarctica primarily in focused, relatively fast-moving flows called ice streams. Understanding the behaviour of these streams is critical for predicting how the Antarctic ice sheets will respond to climate change and how they might alter sea level. It is generally thought that ice streams are lubricated by a mixture of meltwater and fine-grained sediment (known as till) at their base. But

only now have Edward C. King and colleagues taken a good look at how the meltwater is distributed.

Their seismic soundings of the Rutford Ice Stream (pictured) in West Antarctica indicate that beneath the ice there are channels or ‘canals’ of water about 0.4–0.7 m deep, carved into the till. The seismic anomaly that pinpoints the water is about 200 m wide and 1 km long, but the researchers can’t tell yet if this is a single channel or an unresolved series of narrower canals 3–5 m wide. Either way, the liquid surely helps speed the ice flow towards the ocean.

Bill Hapill



LANDSAT

Formulation of a Roman cosmetic

This unguent shares some surprising features with modern moisturizing creams.

The discovery of a small tin canister in London during archaeological excavations of a Roman temple precinct, dated to the middle of the second century AD¹, is a landmark in the study of this class of artefact. Such discoveries from the Roman world are rare² and this is the only one to be found so far with its lid and contents — a whitish medicinal or cosmetic cream — providing a unique opportunity for us to study the ancient formulation.

The canister has an internal diameter of 60 mm and is 52 mm tall. Organic elemental analysis of the slightly granular cream inside (Fig. 1a) reveals that it contains 50% carbon and 8% hydrogen, with no detectable nitrogen or sulphur. This indicates that protein (gelatin, for example) is not a significant component. The cream only partially dissolves in chloroform/methanol (2:1 by volume), and the only soluble components (40% by weight) are fatty acids (Fig. 2a) of animal origin (deduced from the high proportion of saturated $C_{18:0}$ fatty acid).

The odd-number carbon components, which include *iso*- and *anteiso*-isomers, combined with characteristic $\delta^{13}C$ values for $C_{16:0}$ and $C_{18:0}$ (Fig. 2b; ref. 3), are indicative of a ruminant adipose-fat source — for example, cattle or sheep. Analysis of volatile components by head-space gas chromatography with mass spectrometry (GC/MS) revealed no substances that could be attributed to perfume components (such as monoterpenes).



Figure 1 Appearance of the cream. **a**, Roman canister containing the original cream. **b**, A synthetic version prepared according to the formulation obtained after analysis of the material shown in **a**.

Pyrolysis followed by GC/MS of the insoluble residue yielded 1,6-anhydro- β -D-glucopyranose (levoglucosan) and other pyrolysis products characteristic of glucose-based polymers⁴, and the Fourier-transform infrared (FTIR) spectrum matched that of cellulose/starch (ν_{C-O} value was $1,017\text{ cm}^{-1}$; ν_{O-H} value was $3,296\text{ cm}^{-1}$). GC/MS of alditol/acetate derivatives of the hydrolysed residue confirmed that glucose was the only significant monosaccharide⁵; a lack of free glucose indicated the residue's polysaccharide nature. The black/dark-blue colour produced by an iodine test confirmed that the helical moieties (amylose) of starch molecules were well preserved. Quantitative determination of starch and fatty-acid content using FTIR (comparing ν_{max} values of $2,915\text{ cm}^{-1}$ and $1,017\text{ cm}^{-1}$ for fatty acids and carbohydrates, respectively) indicated a mixture of roughly 1:1 by weight.

These analyses account for over 80% by weight of the cream. Gravimetric analysis, involving heating to 850°C , further revealed an inorganic residue that accounts for more than 15% by weight of the material and which was shown by X-ray fluorescence to be dominated by a tin (Sn) compound (results not shown). X-ray diffraction of the cream (Fig. 2c) yielded a scattering pattern, confirming that tin had been intentionally added to the cream as SnO_2 (stannic oxide, or cassiterite)⁶.

The white appearance of the cream suggests a degree of refinement in the technology used to produce it. Animal fat would have been obtained from animal carcasses, and the absence of cholesterol and the presence of $\Delta^{3,5}$ -cholestadiene provide evidence that the fat was heated, possibly with the aim of bleaching it⁷. Starch would have been isolated by treatment of roots or grains with boiling water. White SnO_2 would have been readily produced by heating refined tin metal in air.

We mixed these ingredients together in the correct proportions and also obtained a white cream (Fig. 1b). This cream had a pleasant texture when rubbed into the skin. Although it felt greasy initially, owing to the fat melting as a result of body heat, this was quickly overtaken by the smooth, powdery texture created by the starch: remarkably, starch is still used for this purpose in modern cosmetics. The addition of SnO_2 to the starch/fat base confers a white opacity, which is consistent with the cream being a cosmetic. Fashionable Roman women aspired to a fair complexion, and the Londinium cream may have served as a foundation layer.

White Roman face paint typically comprised lead acetate (or cerussa), prepared by

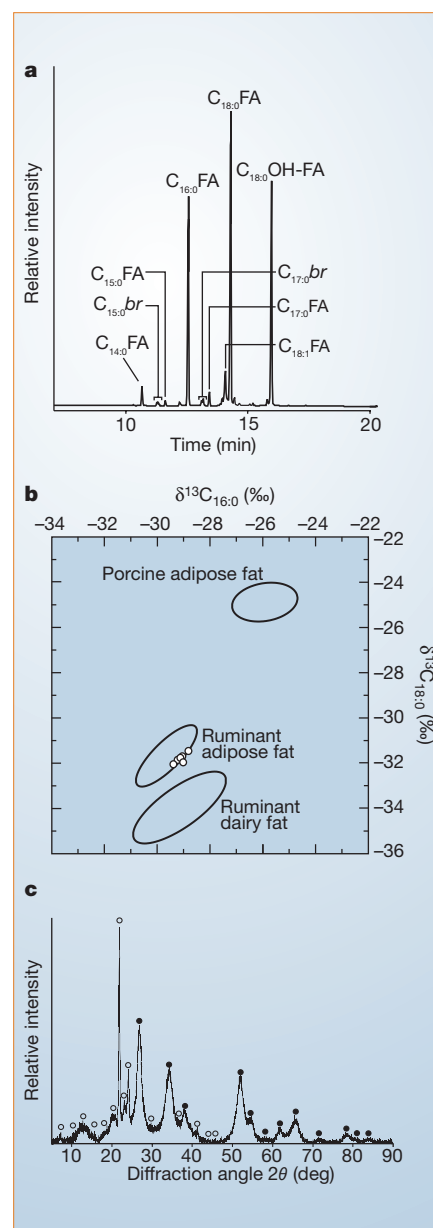


Figure 2 Analysis of the Roman canister's contents. **a**, Partial gas chromatogram of extracted lipids. Peaks: $C_{14:0}$ to $C_{18:0}$, saturated straight-chain fatty acids (FA) with 14 to 18 carbon atoms; $C_{15:br}$ and $C_{17:br}$, *iso*- and *anteiso*-branched-chain fatty acids; $C_{18:1FA}$, mono-unsaturated; $C_{18:0OH-FA}$, hydroxylated form derived by bacterial hydration of the original oleic acid. **b**, $\delta^{13}C$ values for the $C_{16:0}$ and $C_{18:0}$ fatty acids extracted from the Roman cream, compared with confidence ellipses (1σ) corresponding to those from modern cow, sheep and pig adipose fat and sheep and cow butter fat (reference $\delta^{13}C$ values are adjusted for post-Industrial Revolution effects of fossil-fuel burning; analytical precision $\pm 0.3\text{‰}$). **c**, X-ray diffraction pattern of the cream, showing that SnO_2 is present (filled circles). The cystalline fatty acids (open circles) probably formed by deterioration of the animal fat component during burial. Starch is not evident and so must be in a non-crystalline form, consistent with the mode of preparation.

dissolving lead shavings in vinegar⁸. Stannic oxide would have been an acceptable substitute, with supplies being available through the Cornish tin industry. The Romano–British chemists probably believed that they were using a new source of cerussa, so confused were their encyclopaedias in distinguishing lead from tin.

None of the surviving Greco–Roman medical corpora suggest that tin had pharmaceutical value⁹. Although organic tin compounds are now well known for their toxicological properties, inorganic forms seem to be largely inert¹⁰. Hence, unless SnO₂ was added owing to some hitherto unrecognized medicinal attribute, we must conclude that its function was solely as a pigment. The non-toxic properties of SnO₂ would also have been desirable, because by the second century AD the dangers of lead were becoming recognized⁹.

R. P. Evershed*, R. Berstan*, F. Grew†, M. S. Copley*, A. J. H. Charmant*, E. Barham†, H. R. Mottram*, G. Brown‡

*School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK
e-mail: r.p.evershed@bristol.ac.uk

†Museum of London, London Wall, London EC2Y 5HN, UK

‡Pre-Construct Archaeology, Unit 54, Brockley Cross Business Centre, Endwell Road, London SE4 2PD, UK

1. Durrani, N. *Curr. Archaeol.* **192**, 540–547 (2004).
2. Beagrie, N. *Britannia* **20**, 169–191 (1989).
3. Copley, M. S. et al. *Proc. Natl Acad. Sci. USA* **100**, 1524–1529 (2003).
4. Ralph, J. & Hatfield, R. D. J. *Agric. Food Chem.* **39**, 1426–1437 (1991).
5. Blakeney, A. B., Harris, P. J., Henry, R. J. & Stone, B. A. *Carbohydr. Res.* **113**, 291–299 (1983).
6. PDF-2 Powder Diffraction Database (International Centre for Diffraction Data, Pennsylvania, 1999).
7. Cert, A., Lanzón, A., Carelli, A. A., Albi, T. & Amelitti, G. *Food Chem.* **49**, 287–293 (1994).
8. Pliny *Natural History* (ed. Rackham, H.) **9** (Loeb Classical Library, Cambridge, Massachusetts, 1944).
9. Riddle, J. M. in *Dioscorides on Pharmacy and Medicine* 153–154 (Univ. of Texas Press, 1985).
10. *Toxicological Profile for Tin and Tin Compounds* (Agency for Toxic Substances and Disease Registry, US Public Health Service, 2003).

Competing financial interests: declared none.

Biophysics

Water-repellent legs of water striders

Water striders (*Gerris remigis*) have remarkable non-wetting legs that enable them to stand effortlessly and move quickly on water, a feature believed to be due to a surface-tension effect caused by secreted wax^{1–3}. We show here, however, that it is the special hierarchical structure of the legs, which are covered by large numbers of oriented tiny hairs (microsetae) with fine nanogrooves, that is more important in inducing this water resistance.

We used a high-sensitivity balance system to construct force–displacement curves for a water strider's legs when pressing on the

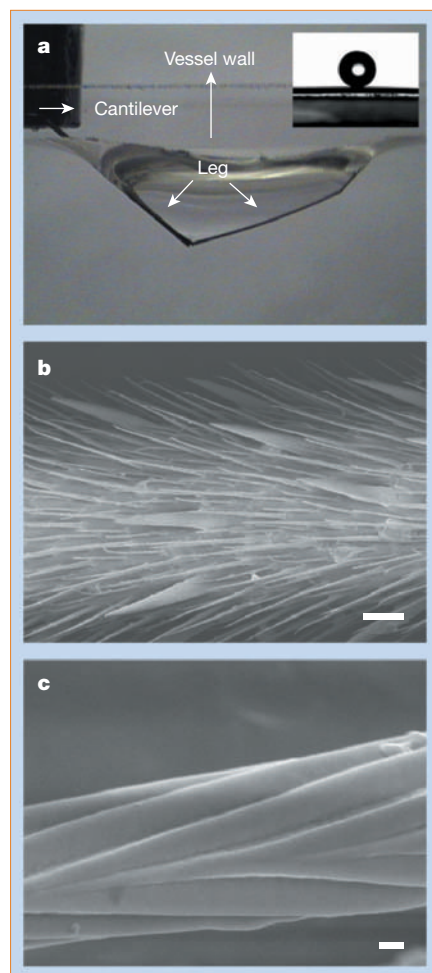


Figure 1 The non-wetting leg of a water strider. **a**, Typical side view of a maximal-depth dimple (4.38 ± 0.02 mm) just before the leg pierces the water surface. Inset, water droplet on a leg; this makes a contact angle of $167.6 \pm 4.4^\circ$. **b**, **c**, Scanning electron microscope images of a leg showing numerous oriented spindly microsetae (**b**) and the fine nanoscale grooved structures on a seta (**c**). Scale bars: **b**, 20 μ m; **c**, 200 nm.

water surface (for methods, see supplementary information). Surprisingly, the leg does not pierce the water surface until a dimple of 4.38 ± 0.02 mm depth is formed (Fig. 1a). The maximal supporting force of a single leg is 152 dynes (see supplementary information), or about 15 times the total body weight of the insect. The corresponding volume of water ejected is roughly 300 times that of the leg itself, indicating that its surface is strikingly water repellent.

For comparison, we made a hydrophobic 'leg' from a smooth quartz fibre that was similar in shape and size to a strider's leg. Its surface was modified by a self-assembling monolayer of low-surface-energy hepta-decafluorodecyltrimethoxysilane (FAS-17), which makes a contact angle of 109° with a water droplet on a flat surface⁴. Water supports the artificial leg with a maximal force of only 19.05 dynes (see supplementary information), which is enough to support the strider at rest but not to enable it to glide or dart around rapidly on the surface.

This finding suggests that the force exerted by the strider's legs could be due to a 'super-hydrophobic' effect (that is, the contact angle with water would be greater than 150°). We verified that this was indeed the case by sessile water-drop measurements, which showed that the contact angle of the insect's legs with water was $167.6 \pm 4.4^\circ$ (Fig. 1a, inset).

The contact angle of water made with the cuticle wax secreted on a strider's leg is about 105° (ref. 5), which is not enough to account for its marked water repellence. Knowing that microstructures on an object with low surface energy can enhance its hydrophobicity⁶, we investigated the physical features of the legs.

Scanning electronic micrographs revealed numerous oriented setae on the legs. These are needle-shaped, with diameters ranging from 3 micrometres down to several hundred nanometres (Fig. 1b). Most setae are roughly 50 μ m in length and arranged at an inclined angle of about 20° from the surface of leg. Many elaborate, nanoscale grooves are evident on each microseta, and these form a unique hierarchical structure (Fig. 1c).

According to Cassie's law for surface wettability, such microstructures can be regarded as heterogeneous surfaces composed of solid and air⁷. The apparent contact angle θ_i of the legs is described by $\cos \theta_i = f_1 \cos \theta_w - f_2$, where f_1 is the area fraction of microsetae with nanogrooves, f_2 is the area fraction of air on the leg surface and θ_w is the contact angle of the secreted wax. Using measured values of θ_i and θ_w , we deduce from the equation that the air fraction between the leg and the water surface corresponds to $f_2 = 96.86\%$. Available air is trapped in spaces in the microsetae and nanogrooves to form a cushion at the leg–water interface that prevents the legs from being wetted.

This unique hierarchical micro- and nanostructuring on the leg's surface therefore seems to be responsible for its water resistance and the strong supporting force. This clever arrangement allows water striders to survive on water even if they are being bombarded by raindrops, when they bounce to avoid being drowned. Our discovery may be helpful in the design of miniature aquatic devices and non-wetting materials.

Xuefeng Gao*†, Lei Jiang*‡

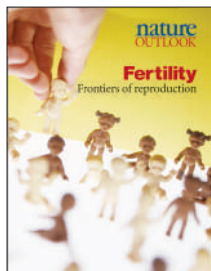
*Key Laboratory of Organic Solids, Institute of Chemistry, and the †Graduate School, Chinese Academy of Sciences, Beijing 100080, China

‡National Center for Nanoscience and Nanotechnology, Beijing 100080, China
e-mail: jianglei@iccas.ac.cn

1. Caponigro, M. A. & Erilsen, C. H. *Am. Midland Nat.* **95**, 268–278 (1976).
2. Dickinson, M. *Nature* **424**, 621–622 (2003).
3. Hu, D. L., Chan, B. & Bush, J. W. M. *Nature* **424**, 663–666 (2003).
4. Tadanaga, K., Katata, N. & Minami, T. *J. Am. Ceram. Soc.* **80**, 1040–1042 (1997).
5. Holdgate, M. W. *J. Exp. Biol.* **32**, 591–617 (1955).
6. Feng, L. et al. *Adv. Mater.* **14**, 1857–1860 (2002).
7. Cassie, A. B. D. & Baxter, S. *Trans. Faraday Soc.* **40**, 546–551 (1944).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.



L. Lawry/SPL

Editor, Nature:
Philip Campbell

Supplement Editors:
Peter Aldhous
Natalie DeWitt
Tim Lincoln
Amanda Tromans

Supplement Publisher:
Sarah Greaves

Subeditors:
Colin Sullivan
Nell Boase

Production Editor:
Alison Hopkins

Art Editor:
Martin Harrison

Diagrams
Alisdair Macdonald
Paul Jackman
Nik Spencer
Radha Clelland

Picture Researcher:
Barbara Izdebska

Production:
Susan Gray

Marketing:
Claire Aspinall

The Macmillan Building,
4 Crinan Street,
London N1 9XW, UK
Tel 44 (0) 20 7833 4000
e-mail: nature@nature.com
www.nature.com/nature



nature publishing group

© 2004 Nature Publishing Group

“All couples and individuals have the basic right to decide freely and responsibly the number and spacing of their children and to have the information, education and means to do so.” Few would disagree with this statement, agreed by world leaders at the United Nations’ International Conference on Population and Development in Cairo, a decade ago.

The arrival of a child is a life-changing event — one that we would all like to control, and plan for. Alas, it often doesn’t work out that way. Infertility curses many couples, whereas lives are turned upside down every day by reproductive ‘accidents’. That’s especially true in the developing world, where access to contraception is limited. But even in the United States about half of all pregnancies are unplanned.

Meeting the UN conference’s goal will not be easy. Religious and moral agendas often complicate the picture, arguing — unrealistically — that unwanted pregnancies are best avoided through sexual abstinence. Economic and social factors weigh heavily: the trend in many rich countries to delay reproduction until later in life helps to explain why many couples then have trouble conceiving. But reproductive empowerment will require a vigorous research agenda, which provides the backdrop to this Outlook.

Mention the science of human reproduction, and thoughts turn to *in vitro* fertilization and similar technologies. In this area, science fact often seems to blur into science fiction. For a glimpse of what might lie ahead, a leading exponent of that genre has devised a brochure for a reproductive services company operating 75 years from now.

Other articles look at the nearer term. One of the most exciting — and controversial — frontiers of fertility is the removal of the menopause. Here, the science is moving fast. So too in other areas, such as how imprinting — the differential expression of genes, depending on whether they come from the mother or the father — controls fetal nutrition.

But many of the field’s frontiers are shrouded in ignorance — and, in some cases, blighted by inaction. Is our biological fertility declining? If so, is environmental pollution to blame? And why are millions of people, even in the developed world, using contraceptive methods that are unreliable and fail to deploy the latest findings in molecular biology?

Our authors don’t claim to have all the answers. But we hope that this Outlook will frame the questions from which a brighter reproductive future can emerge.

Peter Aldhous and Natalie DeWitt



p40

news features

38 The fertility riddle
Declan Butler

40 Age is no barrier...
Kendall Powell

commentary

43 Waiting for the second coming
Jerome F. Strauss III and
Michael Kafrissen

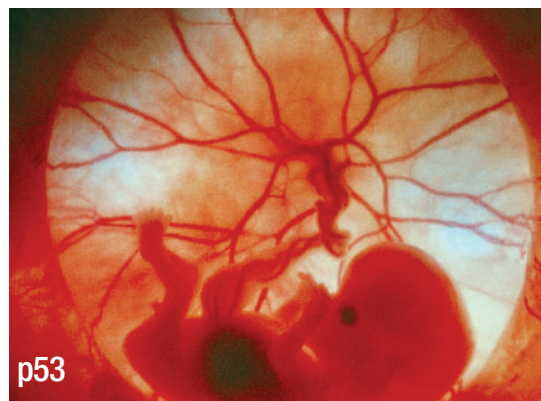
futures

46 Fertility 2079
Greg Bear

news and views features

48 Seeds of concern
R. John Aitken, Peter Koopman and
Sheena E. M. Lewis

53 Resourceful imprinting
Miguel Constância, Gavin Kelsey and
Wolf Reik

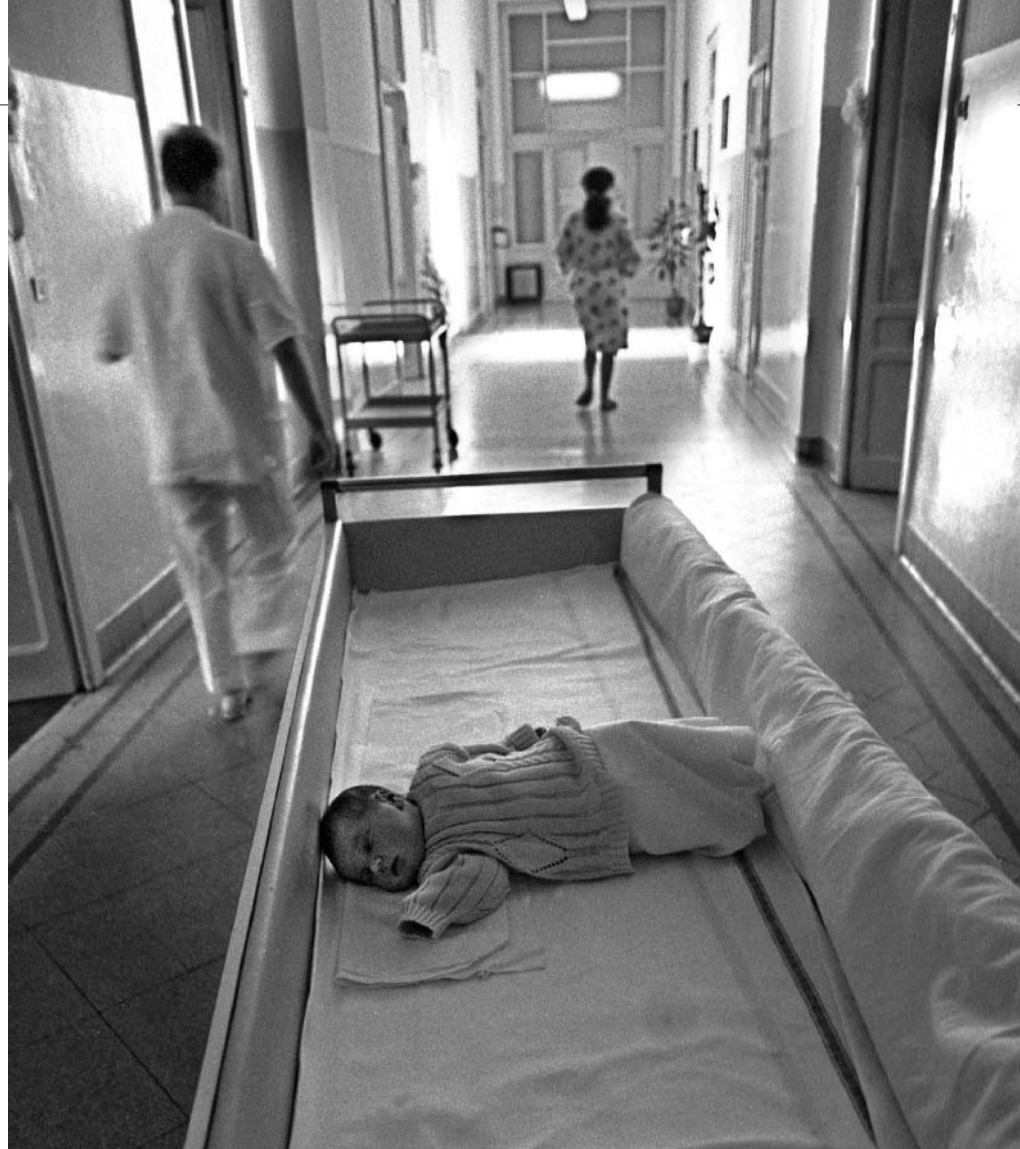


p53

Slow day in an Italian
maternity ward.
Low birth rates
promise future
population decline.

The fertility riddle

Across the developed world, birth rates are plummeting. Is this just a social phenomenon, or is our biological fertility also declining? We don't yet know, and that is worrying, says Declan Butler.



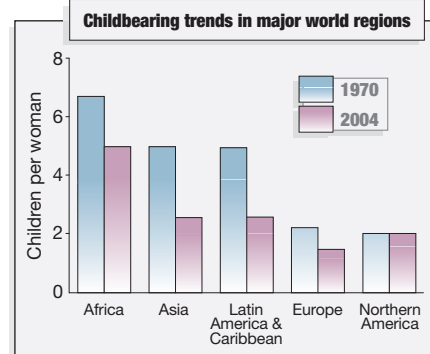
SOURCE: UN

If current trends continue, Japan will be deserted by the middle of the next millennium. The nation's birth rate is so low that its population will peak in 2006 and decline thereafter¹ — raising the prospect of economic chaos as a greying population overwhelms its pension and healthcare systems. It's a similar story in South Korea, Italy, Spain and across most of Eastern Europe. In only four industrialized countries are women, on average, having the two children needed to sustain the population² (see chart).

This demographic change is mostly the result of a social climate in which couples are choosing to have fewer children, or none at all. But might something more sinister be going on, such as environmental pollution or sexually transmitted diseases causing a decline in male or female fertility?

"We must take the possibility seriously, but it's too early to be alarmist," says Henri Leridon, who heads the Laboratory of Epidemiology, Demography, and Social Sciences at the University of Paris XI.

What is alarming is that few credible data have been collected on the issue. If our biological fertility is on the slide, it's happening against a background of scant interest from research organizations. But there are reasons to be concerned. For one thing, women's



fertility declines with age — and this means that the trend to start families later is bound to create problems. Looking ahead, the expansion of *in vitro* fertilization may create a cohort of adults who have inherited their parents' fertility problems.

The shift towards having more sexual partners in a lifetime carries another infertility risk: sexually transmitted diseases. Some 5% of Americans of reproductive age are infected with the bacterium *Chlamydia trachomatis*, a major cause of female infertility — which is rarely diagnosed as it produces no obvious symptoms³. "That's an epidemic," says Peter McGovern, a fertility specialist at the University of Medicine and Dentistry of New Jersey in Newark.

Fertility experts are also concerned that another epidemic — of obesity — may bring infertility. Many severely obese women fail to ovulate, even if they report regular menstrual cycles. And obesity is linked to polycystic ovary syndrome, an important cause of infertility in which ovarian follicles fail to mature. Up to 10% of US women are thought to have this condition; it's hard to be sure of the figure because of disagreements over its diagnosis⁴. "But as it's related to obesity, we are going to see more of it," McGovern predicts.

Born that way

Smoking, alcohol consumption and a range of other lifestyle factors can all reduce a couple's ability to conceive, usually affecting women more severely than men. Most of these factors act through hormonal pathways, so understanding these interactions poses a major challenge to endocrinologists.

But one of researchers' biggest obstacles is that our fertility is mainly determined by the environment we inhabited in the womb⁵. A woman's stock of eggs is defined by the number and maturity of her ovarian follicles when she herself was an embryo. Normal fetal follicular development depends on the mother's diet and other lifestyle factors, including her exposure to chemicals.

Next in line: fertility problems may be passed from one generation to the next.



Adult male sperm count and quality is determined largely by the development of sperm-nurturing Sertoli cells in the embryonic testes. This depends heavily on exposure to sex hormones in the womb, which again is influenced by the mother's lifestyle and other environmental factors.

Prospective parents

So if fertility specialists are to make sense of trends in biological fertility, ideally they should conduct prospective studies relating these environmental factors to the fertility of the next generation in the decades to come.

Until now, most information has come from retrospective studies, in which parents are asked about their lifestyle and environment at the time when they were having children, including their exposure to toxic chemicals. "Trying to measure exposures 30 years ago is very difficult," observes Stewart Irvine, an expert on male infertility at the UK Medical Research Council's Human Reproductive Sciences Unit in Edinburgh.

Such problems have bedevilled attempts to find out whether pollution is reducing sperm counts and sperm quality (see page 48). It has been difficult even to determine whether there really is a decline. Semen samples vary enormously in sperm density and quality, within and between individuals, so you need large sample sizes to detect any drop in sperm counts. But it is hard to recruit volunteers to give sperm, and those who do sign up often have a vested interest: their own fertility problems. "You really can't get representative population samples," complains Michael Joffe, an epidemiologist at Imperial College London, who studies fertility trends.

Women's fertility should be causing as

much concern as men's, but it has attracted less attention. In the United States, a headache for researchers is that medical insurance often does not cover treatment for female infertility. "This means doctors usually code it as endometriosis or fibroids, or something, so that patients can get insurance coverage," says McGovern. "This has a big impact on the data. It's the elephant in the room."

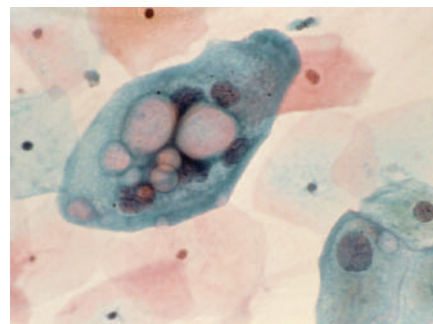
Experts are increasingly using 'time to pregnancy' studies to investigate infertility. In these, they ask couples having unprotected sex how long it took them to achieve a pregnancy, and to relate this information to factors that might influence fertility.

So far, most of these studies have been retrospective, and have raised as many questions as they have answered. Joffe, for example, gave 1,540 Britons aged 16–59 a questionnaire, and found that fertility seems, if anything, to be increasing⁶. The percentage of couples taking longer than one year to get pregnant fell from 21% in 1960–65 to just 10% in 1991–93.

Another study⁷, based on the US National Survey of Family Growth, a survey of thousands of Americans conducted every few years, suggested that fertility declined between the early 1980s and the mid-1990s. The authors attributed this to women starting their families at a later age.

Joffe is preparing to publish a follow-up study of 40,000 subjects. Shanna Swann, a reproductive epidemiologist at the University of Missouri-Columbia, will soon publish a meta-analysis of fertility studies conducted over the past five decades. But Leridon cautions against studies that rely on recalling information about past pregnancies. "These

"Studies need large samples. But it is hard to recruit volunteers to give sperm, and those who do often have a vested interest: their own fertility problems."



Invisible epidemic: Chlamydia infection is a major cause of female infertility.

surveys don't give clear evidence," he says.

What is needed, argue many experts, are studies in which couples are recruited before they conceive⁸. To detect the effects of factors such as environmental pollutants, these studies should record, for instance, frequency of intercourse and timing of ovulation.

Given the difficulty and cost, such studies are few. This has led Roger Short, a reproductive biologist at the University of Melbourne in Australia, to suggest monitoring the incidence of non-identical twins as an alternative strategy⁹. This is a reasonable measure for a combination of male and female fertility, he argues, because it reflects the frequency of double ovulation, the probability of fertilization, and the survival of embryos.

Care would be needed to exclude women taking drugs to stimulate ovulation, and those whose ovulation has been depressed by the prolonged use of oral contraceptives. But it shouldn't be difficult, Short suggests, for researchers to use twins to make inferences about changes in fertility.

Maybe so, but experts aren't betting that this short cut will determine whether our biological fertility is in decline. If governments want the answer to a question that could have profound demographic and economic consequences, argues McGovern, they must provide substantial increases in funding. Ultimately, the science of human fertility is hampered by a chronic lack of research into basic reproductive biology, epidemiology and toxicology — in particular in the developing embryo.

The bottom line, says Irvine, is that studies addressing trends in fertility are "simply very difficult to do. Sometimes I wish I'd chosen to work on something else."

Declan Butler is Nature's European correspondent.

1. *Statistical Handbook of Japan 2004* (Statistics Bureau, MPHPT, Tokyo, 2004); www.stat.go.jp/english/data/handbook/c02cont.htm
2. *World Fertility Report: 2003* (UN Department of Economic and Social Affairs, Population Division, New York, 2004); www.un.org/esa/population/publications/worldfertility/World_Fertility_Report.htm
3. Miller, W. C. et al. *J. Am. Med. Assoc.* **291**, 2229–2236 (2004).
4. Azziz, R. *Reprod. Biomed. Online* **8**, 644–648 (2004).
5. Sharpe, R. M. & Franks, S. *Nature Cell Biol.* **4** (Suppl. 1), S33–S40 (2002).
6. Joffe, M. *Lancet* **355**, 1961–1965 (2000).
7. Chandra, A. & Stephen, E. H. *Fam. Plann. Perspect.* **30**, 34–42 (1998).
8. Tingen, C., Stanford, J. B. & Dunson, D. B. *Env. Health Perspect.* **112**, 87–93 (2004).
9. Tong, S. & Short, R. V. *Hum. Reprod.* **13**, 95–98 (1998).

Age is no barrier...

Advances in reproductive medicine hint that female fertility might be extended into late middle age and beyond. But will the methods be safe? And is society ready for this demographic shift? Kendall Powell investigates.



Benjamin Franklin said that the only certainties in this world are death and taxes. To these, women might add a third: the menopause. But this year, two separate developments have suggested that Franklin was right after all.

In March came a report indicating that ovaries contain stem cells that might be prompted to carry on making eggs throughout a woman's adult life¹. Then, in October, Belgian doctors reported that a woman whose ovaries were damaged by cancer chemotherapy had given birth, possibly owing to a graft of her own preserved ovarian tissue that they had transplanted into her body². Together, these results hinted that women might one day be able to delay starting their families until their forties or beyond.

The scientists working on these projects aim to develop methods to preserve the fertility of women who have been driven through a premature menopause by treatments for diseases such as cancer. But given that many women in industrialized societies want to delay reproduction while they further their careers, such methods are inevitably also going to be seen as tools for healthy women.

At the time of her first menstrual period, a woman's ovaries contain some 250,000 follicles, each of which has the potential to release an egg. But over her reproductive life, a

woman will ovulate no more than 500 times, because her supply of eggs is progressively wiped out in a process called follicular atresia. In her late thirties, a woman's fertility dips sharply because of the declining quality of her remaining eggs and her failure to ovulate during many menstrual cycles. Eventually, by the age of about 50, she has so few follicles left that she can no longer respond to the hormones that stimulate follicular development and ovulation, resulting in the menopause.

The simplest way to extend a woman's reproductive life would be to freeze ovarian tissue or eggs until they are needed. More fundamentally, it might be possible to disrupt the molecular mechanisms behind follicular atresia. Alternatively, the stem cells in a woman's ovaries might be coaxed into replenishing her egg supply.

Cold storage

Although most fertility experts approve of giving women more options, opinions vary on the likely safety and efficacy of methods to delay childbearing. And there's even less consensus on whether society is ready for the consequences of such techniques. Some observers see a troubling future in which children will spend their entire adult lives caring for their elderly parents. But if we are living longer and healthier lives, say others, why restrict reproduction to our twenties

and thirties?

Currently, the only established method of preserving a woman's fertility is to freeze embryos created by *in vitro* fertilization (IVF) for later transfer to the womb. Here, the odds of a successful pregnancy can approach 50% per attempt, for women younger than 35. But this option requires a woman to have a willing partner, or at least a sperm donor, to fertilize her eggs before freezing.

Men can store frozen sperm for later use. But until recently, no one would have suggested that women should freeze their eggs. The mature human egg is the largest cell in the body. As such, it is very sensitive to freezing, because ice crystals can turn its cytoplasm to mush upon thawing.

Recent advances in banking frozen eggs have been spurred in part by the Catholic church's view that embryos must be protected as living individuals. In predominantly Catholic Italy, embryo freezing is frowned upon, and the practice was banned in February this year. Realizing the need for an alternative, Eleonora Porcu and her team at the University of Bologna have developed a liquid — including a sugar and 1,2-propanediol, which acts as an antifreeze — that allows eggs to survive freezing.

Using her solution, and fertilizing eggs by injecting sperm, Porcu has improved the success rate from one birth per 100 frozen



Improvements in freezing eggs and ovarian tissue (below) for reimplantation might one day allow grandparents to have more children of their own.



eggs to about one in five³. “Egg freezing is almost ready to be introduced into the clinical routine,” she concludes. Indeed, the Boston-based firm Extend Fertility already offers egg banking through their US fertility clinics for prices that start at \$12,500.

Banking on a baby

But Porcu and other experts stress that egg banking is not the female equivalent of storing frozen sperm. For a start, getting eggs is not a trivial procedure — the ovaries must be stimulated with hormones and the eggs then harvested with a surgical needle.

The limited number of eggs that can be frozen, compared with the enormous number of sperm in an ejaculate, also means that egg banking is a gamble with future fertility. Any woman who believes that frozen eggs represent “money in the bank” for later fertility is deluding herself, argues Zev Rosenwaks, who directs the Center for Reproductive Medicine and Infertility at Cornell University’s Weill Medical College in New York. “It’s, at best, a 20% shot,” he says.

One way of improving the odds would be to freeze ovarian tissue. The tiny, immature eggs it contains should survive freezing and thawing with few problems. Once back in the body, the graft should begin to churn out eggs. Even if these have to be retrieved manually and fertilized by IVF, the odds of reproducing

might still be reasonably good.

Researchers at Rosenwaks’ Cornell centre, led by reproductive endocrinologist Kutluk Oktay, have been working along these lines. In March, they reported restoring the production of ovarian sex hormones in prematurely menopausal women, whose preserved tissue was transplanted under the skin of their forearm or abdomen.

From one of these patients, Oktay retrieved 20 eggs from the transplanted tissue and found eight that were suitable for IVF with her husband’s sperm. Of these, one egg developed into a healthy-looking four-cell embryo⁴ and was transplanted to the woman’s uterus. Even though Oktay’s procedure did not result in a pregnancy, it was an important proof of principle.

Experts are enthusiastic about applying the technique to women made infertile by cancer treatment, but are much more cautious about its use by healthy women seeking to delay reproduction. “We need to make sure we’re not doing harm to these women,” says Oktay. Given the need for surgery, and the unknown success rate, he suspects that the risks may outweigh the benefits for women with other options for conceiving. The American Society for Reproductive Medicine — a professional organization for fertility clinics — released guidelines on 18 October strongly suggesting that, for now,

egg and ovarian tissue freezing should only be used on an experimental basis for helping chemotherapy patients.

In any case, if researchers can find a safe and subtle way of slowing a woman’s reproductive clock, then the menopause could be delayed without surgery. There are two molecular puzzles that need to be solved if women’s fertility is to be extended in this way. First, researchers need to slow the loss of immature eggs through follicular atresia. And second, they need to maintain eggs’ cellular and genetic quality. Falling egg quality plays an important role in women’s drop in fertility as they approach the menopause.

Stop the clock

Follicular atresia occurs through a mechanism of programmed cell death called apoptosis. At Harvard Medical School in Boston, Jonathan Tilly is trying to slow atresia. In 1999, his team made a mouse lacking a gene called *Bax*, which seems to control the rate of apoptosis. “Their ovaries kept ticking along well into old age, the mouse equivalent of 100 years old,” Tilly says. The older mice could not get pregnant naturally, but through IVF had normal offspring⁵.

But scientists do not know of any drug that could block *Bax* in the body. “Targeting *Bax* is not as easy as it sounds,” Tilly says. So he is now working on a fatty molecule called sphingosine-1-phosphate (S1P), which acts upstream of *Bax* in the signalling pathway that regulates apoptosis. Tilly’s team has shown that ovarian injections of S1P protect mice from a premature menopause caused by radiation^{6,7}.

The problem is that S1P regulates apoptosis throughout the body, and its widespread disruption might have life-threatening consequences. Apoptosis, for instance, is vital in the immune system. So to make S1P a safe way of preserving fertility, Tilly’s team needs to deliver it specifically to the ovaries. They are currently testing a slow-release capsule inserted into the ovaries of rhesus monkeys.

Meanwhile, at Erasmus University in Rotterdam, the Netherlands, researchers led by Axel Themmen believe that they have identified an ovary-specific alternative to S1P. Anti-Müllerian hormone (AMH) directs male embryos to develop male reproductive organs. But women also make AMH, in cells that support egg development in ovarian follicles.

Female mice lacking AMH lose their folli-

J.C. REVY/SPL



cles at three times the normal rate; so Themmen thinks that treatment with the hormone might extend a woman's fertility⁸. But more work must be done first to see the effects of AMH injections on lab animals.

Even staving off follicular atresia will not solve the parallel problem of chromosomal damage in ageing eggs, which is associated with heightened rates of miscarriage and birth defects. Almost 80% of the eggs of women aged 40–45 show abnormal chromosome alignment⁹.

This obstacle would be removed if older women could obtain a fresh batch of eggs. That's where another line of research in Tilly's lab comes in. In a paper published in *Nature* in March, his team presented evidence that adult mouse ovaries contain stem cells that divide to produce new ovarian follicles¹.

Looking more closely at follicular atresia, the team found that follicles were being lost at such a high rate that the supply ought to be exhausted by young adulthood. Second, they observed cells in the ovaries of adolescent mice that looked like the stem cells that exist in fetal females. They also treated female mice with a drug that kills such stem cells, and found that their follicles were rapidly depleted.

If human ovaries contain a similar population of stem cells, it might be possible to avert the menopause by boosting their division. But the idea that female mammals are born with a fixed supply of eggs is well established, and most experts want to see more evidence before accepting Tilly's ideas. "The claims Tilly makes are enormous and revolutionary, but it's been very seldom that one paper overturns a paradigm as deeply entrenched as this one," says Roger Gosden, scientific director of the Jones Institute for

More women are delaying childbirth until later, when fertility begins to decline as developing eggs (below) are lost as a result of apoptosis.



Reproductive Medicine at Eastern Virginia Medical School in Norfolk.

Tilly is tight-lipped about the details of his unpublished work, but feels confident that he will find evidence that women have ovarian stem cells. Oktay's study of his transplant patient also supports this idea. Oktay had originally estimated from the density of follicles in the transplanted tissue that the graft would function for about a year. More than 18 months later, the tissue continues to ovulate regularly — perhaps, Oktay suggests, because stem cells are replenishing the supply of follicles.

Tilly's paper will surely trigger a frenzy of activity as other researchers try to replicate and build from his findings. "The potential

for stem-cell technology applications to reproduction are probably limitless," says Rosenwaks. But Tilly warns against trying to rush into the clinic. "We can't do this faster than we're doing it," he says.

Even so, Tilly's findings have given new urgency to the debate about the social implications of delaying menopause. Will older mothers be able to give their children the support that they need? Is this a fair question, given that elderly men can already become fathers? What are the implications for the demographics of society as a whole? Will it lead to a decline in fertility at the population level, as women eschew reproduction when they are most likely to conceive?

On the last point, reproductive biologists admit to concern. "Natural reproduction is still more efficient than anything we can do in the clinic," says Rosenwaks. "We should educate the public that the earlier, the better your chances, and the more likely you will have a healthy child."

Most fertility experts are unwilling to speculate on the wider consequences of extending women's reproductive lifespan. They argue that decisions about reproduction are a matter for individuals, and say that there should be no absolute upper age limit for motherhood. "I'm against any fixed line," says Gosden.

Outside of reproductive medicine, academics who have considered the issue wonder whether the focus on biological solutions is misguided. If the workplace was organized in ways that allowed women to have children in their twenties or early thirties without compromising their careers, they say, then the pressure to delay childbearing would be relaxed. "We're thinking about stopping the biological clock here, but why not work with the career clocks?" asks Claudia Goldin, an economist who studies women's employment issues at Harvard University. ■

Kendall Powell is a science writer based in Colorado.

1. Johnson, J., Canning, J., Kaneko, T., Pru, J. K. & Tilly, J. L. *Nature* **428**, 145–150 (2004).
2. Donnez, J. et al. *Lancet* **364**, 1405–1410 (2004).
3. Porcu, E. et al. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **113** (Suppl. 1), S14–S16 (2004).
4. Oktay, K. et al. *Lancet* **363**, 837–840 (2004).
5. Perez, G. I. et al. *Nature Genet.* **21**, 200–203 (1999).
6. Morita, Y. et al. *Nature Med.* **6**, 1109–1114 (2000).
7. Paris, F. et al. *Nature Med.* **8**, 901–902 (2002).
8. Gruijters, M. J., Visser, J. A., Durlinger, A. L. L. & Themmen, A. P. N. *Mol. Cell. Endocrinol.* **211**, 85–90 (2003).
9. Battaglia, D. E., Goodwin, P., Klein, N. A. & Soules, M. R. *Hum. Reprod.* **11**, 2217–2222 (1996).

Waiting for the second coming

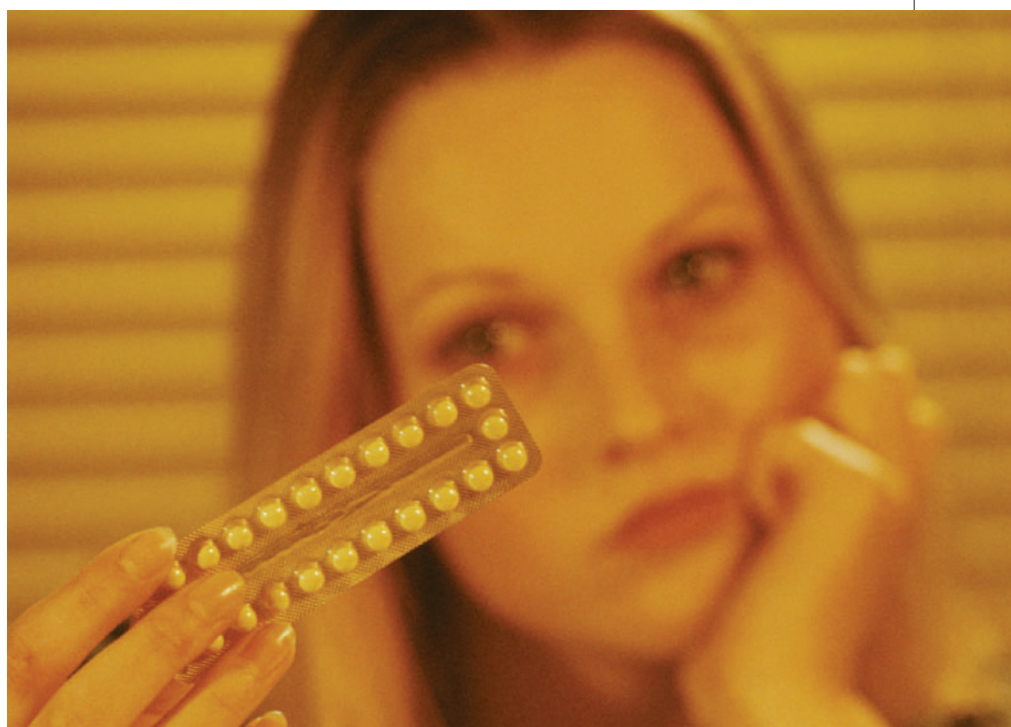
Contraceptive research is seriously in need of revitalization.

Jerome F. Strauss III and
Michael Kafrissen

Even developed countries have a staggeringly high incidence of unplanned pregnancies. In the United States about half are unintended at the time of conception, and many involve couples practising some form of contraception. This reflects the inconsistent use of effective methods and a discontinuation rate approaching 50% after one year. Contraceptives are simply not meeting the needs of society. Despite epidemiological and demographic data that make a compelling case for immediate and substantial investment in contraceptive research and development, the academic community and industry seem unmoved by these needs.

Successive reports issued over more than a decade by the Institute of Medicine (IOM) show that a second contraceptive revolution, or at least a revitalization of contraception research, is sorely needed^{1–3}. In developed countries, the high discontinuation rate shows that consumers want safer, more effective and more user-friendly contraceptives. In contrast, the developing world lacks universal access to effective, inexpensive and appropriate contraception — let alone methods that also protect against sexually transmitted diseases. The widely used oral contraceptives have adverse cardiovascular side effects in a small number of users, and there is debate about a possible small cancer risk. And apart from vasectomy and condoms, there are virtually no male methods.

Yet the number of patents relating to contraception is modest compared with the number issued for drugs to treat high blood pressure (Fig. 1). Only 2.3% of the new molecular entities approved by the US Food and Drug Administration (FDA) during the past six years have been for contraceptive products, and all of these were steroids, and so only incremental advances (Fig. 2). There are more than 1.5 billion women of reproductive age, more than one billion of whom are married and presumably in need of some family-planning method, whereas an estimated 600 million people have high blood pressure.



Pregnant pause: the discontinuation rate for contraception is almost 50% after one year.

Given the universal need, fertility regulation should rank among the most important quality-of-life concerns. So what gives?

This is a problem of the compartmentalization of reproductive sciences and clinical development, and a failure to direct resources, both monetary and legislative, into programmes that will turn reproductive biology into products. Unfortunately, most reproductive biologists have little knowledge of drug development and do not appreciate what makes an attractive drug target, even at the very early stages of the process. And academia contains very few translational scientists who can advise on product

development and early-stage clinical trials. Here we examine some of the factors impeding progress in contraceptive development, and offer some suggestions to address them.

Lost in translation

Two obvious places to seek targets for contraceptive development are the genes and proteins selectively expressed in the female and male reproductive organs. Reproduction is all about interactions between cells, changes in membrane structure and membrane fusion, which are critical to processes such as the preparation of sperm for fertilization and fertilization itself. Carbohydrate moieties and lipids are crucial to these processes, and the reproductive tract, particularly of the male, contains unique forms of them. We hope that forays into 'glycomics' and 'lipidomics', new technologies for large-scale analysis of carbohydrates and lipids, will reveal promising new leads.

Although this is a good starting place, there is a bottleneck between basic science and product development. Academic investigators tend not to be attracted to applied research, as they are uncertain how funding agencies will receive it. Few of them have access to the costly infrastructure, such as high-throughput screening facilities and chemical libraries, needed to move on to the next step.

To feed the pipeline of potential contraceptives, academic investigators will need to better understand how the system works, how to access the tools needed to move their discoveries along, and how to 'de-risk' a concept, target or lead compound to make it tantalizing enough for industry to bite. This may require new data to overcome concerns

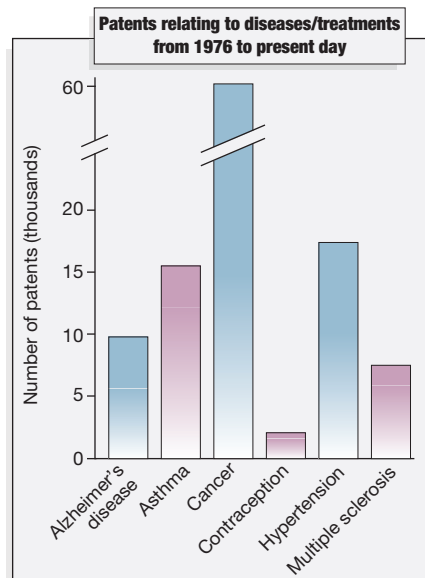


Figure 1 Number of US patents issued for contraceptives and other treatments.

GETTY IMAGES

SOURCE: US PATENT & TRADEMARK OFFICE

Five years of fighting litigation over the safety of the hormonal implant Norplant, although successful, was a heavy financial burden for its manufacturer Wyeth.



about efficacy, delivery method, cost of production and market penetration.

Academic scientists might also need a reality check. RNA interference may be a great contraceptive for worms, but its near-term potential in human contraception is questionable. Ideas for post-fertilization contraceptive methods are nixed by some from the beginning because of a potential backlash from pro-life advocates. Consideration of the social perspective to answer the all-important question, 'If we build it, will they use it?', is rarely contemplated by the bench and translation investigators in academe.

Out in the cold

Another obstacle is the paucity of researchers focused on contraception. There is no shortage of reproductive biologists. The Society for the Study of Reproduction, for example, has about 2,300 members, yet only a few identify contraception as their primary interest. The society's annual meeting this year did not have a single session dedicated to contraception. Basic reproductive biology is not synonymous with contraception research; there is a difference between an investigator who studies the process of ovulation and one trying to discover how to inhibit ovulation.

Contraception research needs to attract talented scientists and physician-investigators who can offer advice on product development and clinical trials. It is not clear whether the National Institutes of Health's Loan Repayment Program (LRP), which aims to attract such workers into contraception and infertility research by paying for debts accumulated during undergraduate and graduate medical education, has addressed these needs, at least for contraception. For example, of 56 individuals originally receiving LRP support, only five were doctoral-level investigators whose research was directly relevant to contraception. Few existing graduate and postgraduate programmes focus on contraception, and those that do tend to emphasize epidemiological studies and late-stage clinical trials.

It is rare for reproductive biologists to focus on the nitty-gritty work of getting a new target or compound off the ground. This problem can only be resolved by federal funding agencies and foundations establishing infrastructure and training programmes to cultivate such individuals.

Healthcare industries are essential for

developing contraceptive products, but there is no guarantee of interested parties. A number of questions must be addressed. For example, how well does a contraceptive fit with the current business and customer base? What advantages does it offer over existing methods? How difficult will it be to carry out the necessary research, what is the likelihood of approval, and what are the liability risks? What are the competing products and the likely returns on the investment? Without a strong business rationale and a champion to propel a development plan forward, contraception may be a non-starter in many companies.

Industry research has a finite budget, and most companies have a stable of competing therapeutic needs and business cases. Investment decisions must be weighed carefully, given that only 22% of phase I candidates are ever approved, and there is only a 57% chance of drugs surviving phase III clinical trials and winning FDA approval. The gamble on contraception R&D will be most attractive to companies already involved in some related field, where they can just expand their core competencies.

Showing that a new contraceptive is as safe or safer than current products could take many years — possibly longer than the patent life. Developing better products requires scientific validation, and the affordability of this is an ever-present issue. It may be worth targeting development at a particular segment of the population, such as men at risk of sexually transmitted diseases, women at different times after childbirth, poor compliers or diabetics. But such segmentation of the user base may magnify the previously mentioned challenges, and limit the financial returns.

The approval process for women's contraceptives is relatively straightforward;

there is much less experience for men's products. Existing approaches, such as the steroid hormones oestrogen and progestin, are well served by current preclinical and clinical guidelines, but it is not clear that these will suffice for novel methods, or whether compliance with these principles will satisfy the tort system.

Some insurance companies' refusal to provide liability coverage for contraceptives further hampers new product development. Contraceptive methods are used primarily by healthy populations, potentially for long periods of time. This would seem to encourage drug development, but it is counterbalanced by the risk of making healthy people ill. Given the large number of potential users, even a small proportion of adverse effects would have dire consequences, not only for the manufacturer, but also for public perception of the safety and value of specific methods or contraception in general. Unfortunately, without national or international databases there is little

hope that reliable epidemiological data could inform regulatory and legal solutions.

The chequered past of contraceptive development has made the healthcare industry cautious. For example, plaintiff attorneys — perhaps inspired by the litigation surrounding

silicone breast implants — targeted Wyeth over its contraceptive Norplant, a hormonal preparation encapsulated in a silicone elastomer. This was despite the FDA having no concerns over the drug's safety. During more than five years in court, the manufacturer won three jury verdicts, 20 pretrial judgments and dismissal of some 14,000 claims. A trial verdict against the manufacturer was overturned on appeal. Despite the favourable rulings, defending the product was a significant financial burden for Wyeth.

It should be no surprise, then, that the president of a large pharmaceutical research programme once told one of us that they would not touch contraception with a ten-foot pole.

A modest proposal

We have briefly touched on the opportunities and obstacles to contraception research, specifically those bottlenecks that limit collaboration between academia and industry. There are a number of under-explored

"There is a difference between an investigator who studies the process of ovulation and one trying to discover how to inhibit ovulation."

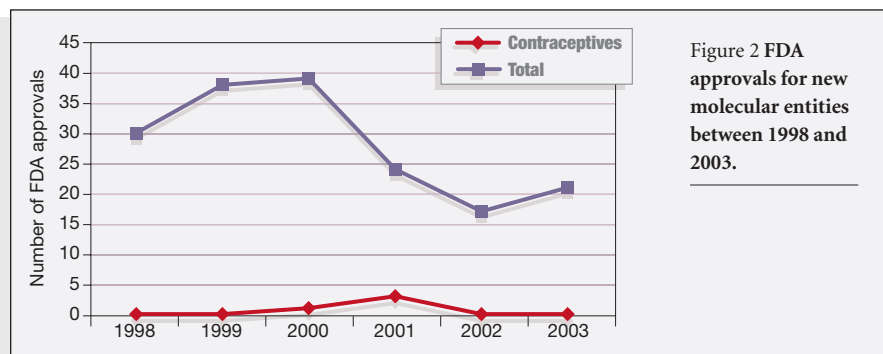


Figure 2 FDA approvals for new molecular entities between 1998 and 2003.

SOURCE: FDA

opportunities as well as measures to cross hurdles, but none alone will be sufficient.

First, there needs to be better interaction between academic investigators and the healthcare industry and regulatory agencies. These discussions should try to understand the impediments to product development. The notion of a round table of stakeholders, as proposed in the 2004 IOM report, resonates well with us³. Expanded partnerships between regulatory and academic bodies may foster forward-looking guidelines for novel methods.

Second, targeted public or private funding, perhaps from foundations concerned with global health, is needed to bridge the gap between discovery and product development. Those controlling funding must understand all aspects of contraceptive development. The US federal budget for extramural and intramural programmes related to translational contraception research needs to be increased, and the funds committed should be restricted to research targeted towards contraception, not tangential science. Cooperative funding mechanisms, in which NIH staff participate with extramural investigators, can make this possible, but NIH staff need the funds and technical expertise to invest in promising research. The Consortium for Industrial Collaboration in Contraceptive Research was an attempt to deal with some of these issues, and lessons from its successes and failures should be heeded.

The new NIH Roadmap offers opportunities to build the multidisciplinary research teams needed to pursue contraception research, if committed investigators can pool their strengths to address the translational hurdle. A recent request for applications from the United States Agency for International Development (USAID), which could provide \$125 million over five years for research in contraceptive and reproductive technologies, is one encouraging sign of continuation of investment in contraception research at a time when a number of founda-

tions previously active in this arena (Ford, Andrew W. Mellon, Rockefeller) have redirected their resources elsewhere.

Third, to stimulate industry interest in contraception, the nature and duration of trials and surveillance should be clarified. This would allow an estimate of the investment needed, which is especially problematic for methods targeting both sexually transmitted infections and contraception, as the efficacy of each activity must be evaluated.

Additional incentives for investment in contraception research are also needed. These could include tax benefits, expanded intellectual-property protection, and giving contraceptives the same status as vaccines — as public-health drugs. To address the thorny problem of liability insurance, we should consider: creating an independent review process to set compensation and oversee compensation programmes; having these principles of science and law set by experts rather than lay juries or plaintiff lawyers; establishing alternative compensation models for plaintiffs funded by industry, such as a compensation fund built on industry contributions and dispersed according to expert opinion, not jury award; and expanding liability protection for companies when the FDA has approved a contraceptive.

Fourth, there is an urgent need to support training, career development and retention in contraception research. This can in part be addressed by training grants and mentored career development awards. Neglected areas such as male reproductive health deserve special attention. But despite the good news about the USAID request for applications, relatively few organizations could compete for this funding, limiting the number of sites for training and career development. This can be remedied by increasing the number of investigators in contraception research. Some of these individuals might become

champions of contraceptive research in the healthcare industries.

Fifth, and finally, industry and foundations should seek opportunities for product development beyond the United States and Europe. For example, the biotechnology industry has flourished in Cuba, where it ranks third in economic terms behind sugar and tourism. Brazil has a long history of contributions to contraception research and a flourishing biotechnology sector. These successes, and those in other countries with smaller biomedical-research enterprises, reveal the promise of science in the developing world to contribute to contraception.

The development of methods that let humans promote and suppress fertility is one of the greatest scientific achievements of the past 50 years. But family-planning problems are far from solved, and the sense of mission and urgency that spurred the development of the first widely used contraceptives in the late 1950s and early 1960s is

“Given the large number of potential users, even a small proportion of adverse effects for a new contraceptive would have dire consequences for the manufacturer and for public perception.”

largely missing. Unfortunately, history tells us that the fervour may not emerge from academia or industry. Recommendations made in past IOM reports and those put forward here may languish without a missing ingredient. The first contraceptive revolution was sparked, and partly

sponsored, by visionary lay people such as Margaret Sanger and Katherine Dexter McCormick. A concerned public with faith in science's ability to improve reproductive health is a powerful motivating force and an essential ally in the campaign to catalyse research in contraception.

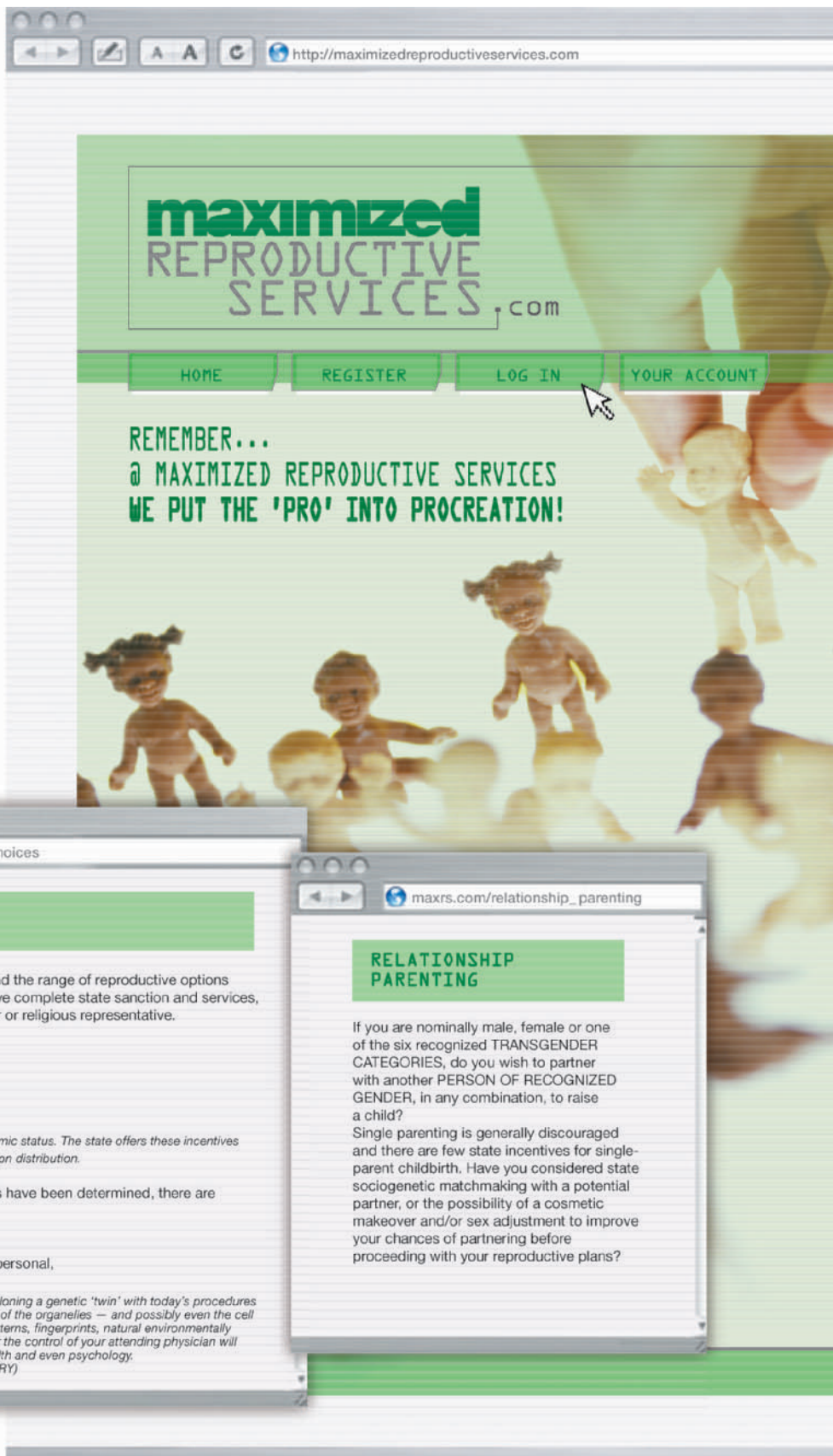
Jerome F. Strauss III is director of the Center for Research on Reproduction & Women's Health, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6142, USA.

Michael Kafrissen is vice-president, clinical affairs, at Ortho-McNeil Pharmaceutical, Route 202 South, Raritan, New Jersey 08869, USA.

1. Mastroianni, L. Jr, Donaldson, P. J. & Kane, T. T. (eds) *Developing New Contraceptives* (National Academy Press, Washington DC, 1990).
2. Harrison, P. F. & Rosenfield, A. (eds) *Contraceptive Research and Development — Looking to the Future* (National Academy Press, Washington DC, 1996).
3. Nass, S. J. & Strauss, J. F. III (eds) *New Frontiers in Contraceptive Research* (National Academy Press, Washington DC, 2004).

Fertility 2079

Greg Bear glimpses the future of human reproduction.



WISE CHOICES

Not only is reproduction essential, it can be safe and fun! And the range of reproductive options available to a modern adult is astonishing. But first, to receive complete state sanction and services, you must describe your goals to a social services counsellor or religious representative.

WHY SHOULD I REPRODUCE?

- Personal enhancement
- Perpetuation of a family name or position
- Religious imperative
- Population replacement and social stability*

**State incentives can be applied for based on genetic, social or economic status. The state offers these incentives as part of society's compelling interest in maintaining a stable population distribution.*

Once your personal, financial, religious and emotional needs have been determined, there are many biomedical options:

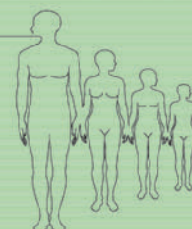
- Natural insemination and legacy ('status') childbirth
- Artificial insemination
- Any of a number of cloning* options followed by implant/personal, implant/surrogate, or ex-utero (cyberwomb) gestations.

**Cloning is often confused with the copying of an adult human. Even cloning a genetic 'twin' with today's procedures will not result in an adult identical to you unless you also layer in many of the organelles — and possibly even the cell membranes — that you received from your mother. Even then, hair patterns, fingerprints, natural environmentally induced recombination within tissues and epigenetic factors not under the control of your attending physician will also be different. This can lead to diverse outcomes in physiology, health and even psychology. (See section 4-C, CLONING A PERSONAL ORGAN BANK REPOSITORY)*

RELATIONSHIP PARENTING

If you are nominally male, female or one of the six recognized TRANSGENDER CATEGORIES, do you wish to partner with another PERSON OF RECOGNIZED GENDER, in any combination, to raise a child?

Single parenting is generally discouraged and there are few state incentives for single-parent childbirth. Have you considered state sociogenetic matchmaking with a potential partner, or the possibility of a cosmetic makeover and/or sex adjustment to improve your chances of partnering before proceeding with your reproductive plans?



> FOR PROSPECTIVE PARENTS >> 2079 NEW EDITION | GREG BEAR

VIEW BASKET

CHECKOUT



WHAT ABOUT BABY?



WISE CHOICES



RELATIONSHIP PARENTING



WE'RE HERE FOR YOU!



THE NEXT BIG STEP

Most couples today take advantage of a surprising number of medical and genetic advancements to help guarantee total success in reproduction. But a word of warning. If you are still considering a purely natural sexual and reproductive experience, think again: 'natural' is often the same as 'dangerous'. Difficulties during pregnancy and development are not uncommon. And the financial burden in today's PERSONAL RESPONSIBILITY CULTURE could be extreme! Can you GUARANTEE to your partner, and your child, that legacy ('status') childrearing will be a positive experience? (See section 11-D, HOW TO WORK WITH YOUR REPRODUCTIVE GUARANTOR AND BONDING COMPANY IN A TIME OF CHANGE)

THE TWO OF YOU, TOGETHER...

If you still plan on engaging in legacy ('status') parenting, are you looking forward to natural childbirth? Or will you subscribe to an alternative process? Please consult your reproduction counsellor on which procedures are approved for your chosen religious or sociopolitical category.

PARENTS IN CRISIS

Do you wish your child to be raised and educated for you? Options are available and can be remarkably reasonable. Most parental rights are retained. Consult your state crisis-aid counsellor for further details.

THE FUTURE IS THEIRS

What social and employment expectations do you have for your offspring? Consider law enforcement preconditioning for your offspring. (See section 47-S, GUARANTEED EMPLOYMENT FOR YOUR PRE-DESTINED CHILD!)

LOVING LEGACY

For
you
tech
read
cras



WE'RE HERE FOR YOU!

FEN

Peo
do
chie
of is

It is the goal of all reproductive service providers to ensure that your childbearing and childrearing experiences are overwhelmingly positive and beneficial for you, for those around you, and for society as a whole. Personal choice and freedom (and spontaneity!) certainly play a role in making your final decision. But centuries of experience have allowed professionals to design and even mandate programmes that reduce the emotional or physical pain and failure rates typically associated with random mating behaviour and bad partnering choices.

[COPYRIGHTED 2079 MAXIMIZED REPRODUCTIVE SERVICES® | DESIGNED BY ALISDAIR MACDONALD]

Seeds of concern

During the past few decades, worries about environmental threats to human health have centred on the possible induction of cancers. Now risks to the male germ line, both real and potential, are also causing disquiet.

R. John Aitken, Peter Koopman and Sheena E. M. Lewis

What do male alligators in Florida and male industrial workers in California have in common? The answer is that, in the latter part of the twentieth century, both provided landmark case histories showing the severe effects that pesticides can have on fertility. Since then investigations of the adverse influence of 'xenobiotics' — molecules that are foreign to biological systems — on male reproduction have turned up more evidence, of various kinds, that all is not well in the man's world.

During the past 50 years, the rapid expansion of the chemicals industry in both the developed and developing worlds has resulted in the release of a plethora of xenobiotics into the environment^{1,2}. These alien molecules have worked their way into our lives in a variety of forms, including pesticides, herbicides, cosmetics, preservatives, cleaning materials, municipal and private waste, pharmaceuticals and industrial by-products. Awareness of the biological risks of chemical toxicity has increased considerably in recent years, but some of these chemicals have long half-lives and have been detected in environmental samples 10–20 years after they were banned for sale or use.

Analysis of the biological fallout from environmental pollution has generally centred on the risks for induction of certain kinds of cancer. But it is becoming increasingly apparent that another major target of

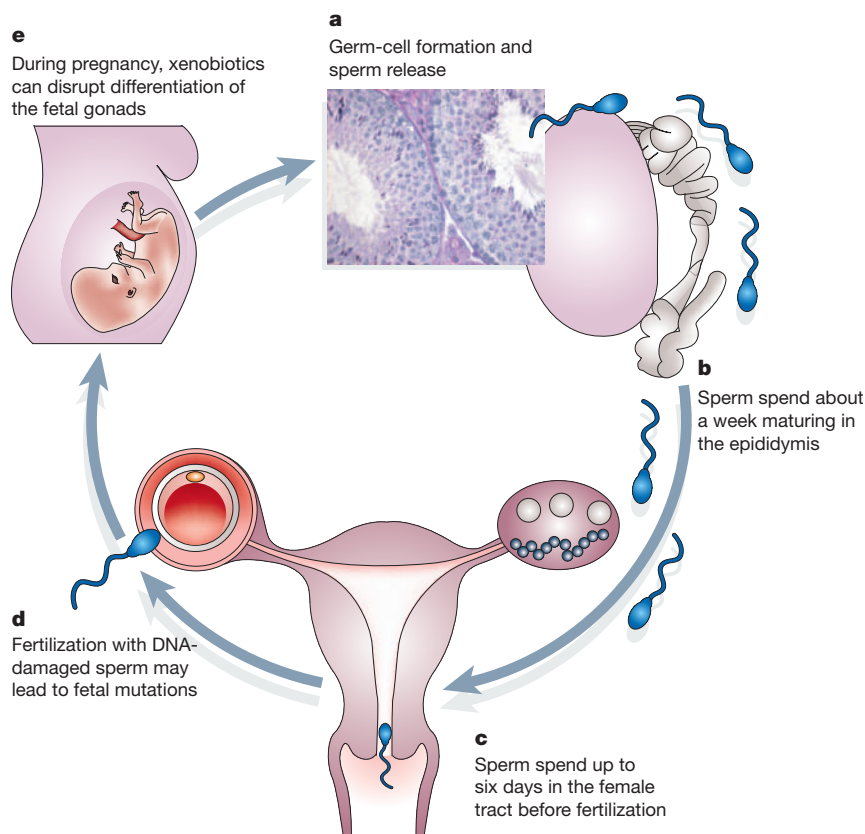


Figure 1 Points in the life cycle of male germ cells when they are vulnerable to xenobiotics. **a**, During their formation in the germinal epithelium, germ cells lose their capacity to self-destruct in response to injury, or to repair damaged DNA. They also lose the protection afforded by Sertoli cells. **b**, **c**, Following their release as sperm, they become susceptible to damage and remain so during their perilous journey through the male epididymis and the female reproductive tract. **d**, DNA damage carried into the egg by a fertilizing sperm must be repaired; any deficiencies in this process will create mutations that may affect the wellbeing of the fetus and offspring. **e**, Later, during pregnancy, xenobiotics can disrupt the differentiation of both germ and non-germ (somatic) cells in the fetal gonads.

this chemical barrage is the reproductive system, particularly in the male². This was first recognized more than 30 years ago, when male workers exposed to 1,2-dibromo-3-chloropropane, an agricultural control agent used to kill nematodes, exhibited severe disruption of sperm development and infertility³. Since then, numerous independent studies⁴ have associated occupational exposure to pesticides, herbicides, industrial agents and heavy metals with poor semen quality and impaired fertility. Although male reproduction can be affected by a variety of

mechanisms that affect hormone balance and other metabolic systems, the disruption of germ-cell differentiation and sperm quality seems to involve two fundamentally different routes of exposure (Fig. 1a–d, and Fig. 1e).

First, xenobiotics and other environmental factors such as radiation can act directly on male germ cells within the mature testis. The highly effective proofreading and repair of DNA in the stem cells that produce sperm means that the male germ line has one of the lowest spontaneous mutation rates in the body⁵. But as these cells go through meiosis,

the cell-division process that produces reproductive gametes, their capacity for DNA repair is reduced and their ability to respond to such damage by undergoing programmed cell death is progressively lost. Moreover, once they are released from the tissue that produces them, the germinal epithelium, male germ cells can no longer rely on the protection previously afforded by their nurse cells in the testes, the Sertoli cells.

Thus, as soon as sperm are released from the germinal epithelium, they are on their own. Bereft of the cytoplasm that houses protective enzymes such as catalase or superoxide dismutase in somatic (non-germ) cells, sperm are committed to a sojourn of about a week in the epididymis, the duct system in which they are remodelled in preparation for ejaculation. Subsequently, they must spend up to six days swimming around the female reproductive tract searching for an egg. During this long and perilous marathon, sperm are particularly vulnerable to DNA damage by a variety of environmental factors⁶. All in all, the sperm is much more susceptible to damage than the egg because of its prolonged solitary existence and relative lack of protective, repair and self-destruct mechanisms.

The second route by which xenobiotics exert an influence on male reproduction is less direct, through exposure of women during pregnancy and subsequent disruption of reproductive tract development in male embryos (Fig. 1e). Such action is thought to affect both the germ cells and the somatic tissues of the male tract, and the consequences include a complex array of pathological changes collectively known as the testicular dysgenesis syndrome, or TDS, in the offspring. The features of TDS include poor semen quality, hypospadias (defective development of the urinary tract), testicular cancer and cryptorchidism (the failure of one or both testes to descend).

The various symptoms of TDS have common risk factors, such as low birth weight, retained placenta and previous pregnancy history, supporting the idea that they have a common cause involving the perturbation of normal fetal development⁷. The importance of abnormal gonad development in testicular cancer is also supported by analysis of the seemingly unaffected testis of men with this condition. Although the gross testis anatomy is apparently normal,

Box 1 Uncertain science and male fertility

Analysing environmental effects on male reproduction is a highly complex business. For example, measuring male fertility is a tough proposition in itself (see page 38). The correlations between fertility and measures of semen quality such as sperm count are weak, and are complicated by factors that include the relative fertility of the female partner and the inherent variability of human semen (which is influenced by abstinence, stress, disease and even the time of year). Moreover, there are selection biases in obtaining semen samples.

Epidemiological studies are dogged by various problems. We do not understand how chemicals are metabolized within the

male reproductive tract, and measuring the type and duration of exposure to xenobiotic agents is fraught with error (self-reporting surveys are a very blunt instrument in this context). There are also great difficulties in establishing relationships between pathological conditions seen in children or young adults and parental exposure to potential toxicants. Long periods of time may have elapsed between parental exposure and the appearance of a condition such as offspring infertility, and distinguishing between maternal and paternal exposures may be impossible where environmental factors are concerned.

Finally, estimating DNA damage in male germ cells is an inexact science. Commonly used techniques (such as the sperm chromatin structure assay and the so-called Comet test) give a general indication of damage, but are only semi-quantitative and provide no information on the exact nature or cause of the damage observed.

All in all, the evidence for and against toxic environmental substances as key factors in male infertility and testicular cancer is nowhere near as clear-cut as we would like. Nonetheless, such evidence as we can trust indicates that here is a serious public health issue requiring closer scrutiny. ■

a closer inspection reveals disordered development of all of the major cell types within the testes, including Sertoli cells, the testosterone-secreting Leydig cells and the germ cells themselves. Experimental evidence for the xenobiotic induction of TDS comes from administration of a testicular toxicant (dibutyl phthalate) to pregnant rats, which produces TDS-like tissue abnormalities in the testes of male offspring⁸. If xenobiotics are involved in causing TDS, they must act relatively early in fetal development. Testicular germ-cell tumours (the most common cancer in young men aged 15 to 35 in Western countries) develop from a precursor condition, known as carcinoma *in situ*, that derives from the earliest stages of germ-cell development in the fetal testis.

Trends in male reproduction

Whatever the route of exposure, environmental factors can clearly affect the development and function of the male reproductive tract. This was first recognized in animal species, a famous example being the disordered reproductive development of male alligators in Lake Apopka, central Florida, following an insecticide spill in the early 1980s^{9,10}. The alligators had abnormal

differentiation of male reproductive organs and lowered levels of testosterone, the steroid hormone that is central to male reproductive biology. Similar findings have been reported for fish, initially in England and more recently in the rivers of other Northern European countries. A mixture of compounds originating from the environmental degradation of certain industrial and household detergents, as well as the urinary excretion of metabolites originating from the female oral contraceptive, seem to be responsible for these emasculating effects on aquatic species¹¹.

In addition, there have been several claims of a deterioration of semen quality in the human male. This issue took wing in 1992, with the publication¹² of a meta-analysis of the biomedical literature that reported on sperm counts. Within this data set, the authors detected an approximate halving in the concentration of sperm in human ejaculates between 1940 and 1990. Analysis of larger data sets supported these general trends and suggested rates of decline in Europe and Australia (3% per yr) that were higher than those observed in either the United States (1.5% per yr) or non-Western countries, where no decline was seen —

albeit on the basis of limited data¹³. Additional studies from Edinburgh¹⁴ indicated that the secular trend in semen quality is a 'birth cohort' effect — that is, the age of the subject when semen analysis is performed does not matter; rather it is the date of birth that is important. According to these data, men born before 1959 have significantly higher motile sperm counts than men born after 1970. So it is not just wisdom and experience that distinguishes the lecturer from his students.

The situation seems to be particularly severe in Denmark, where low fertility rates have been linked with poor semen quality: 25% of 19-year-old Danish men currently exhibit sperm counts in the subfertile range (less than 20 million per ml)¹⁵. But not everyone is convinced by these data on falling sperm counts, and several studies have failed to confirm these trends¹⁶. Certainly, the difficulties in securing reliable data in this area are considerable (Box 1).

If semen analysis were the only method we had of measuring male reproductive potential, the possibility of falling sperm counts would be more a cause for curiosity than concern. However, global trends in testicular cancer bear out the view that the male reproductive system is under attack: the incidence of such cancer has increased in Caucasian men in all developed countries; the current lifetime risk is 0.3–0.8%¹⁷. This contrasts with cancers of the female reproductive tract, with risks that remained largely unchanged or actually declined during the same period (Fig. 2). The increase in testicular cancer cannot be accounted for by the fact that we are living longer or have better methods of detection. Testicular cancer is a disease of young men and is easily detected, but its rising incidence is certainly a cause for concern, even if it is still a relatively rare condition.

Indications of a possible link between low sperm counts and testicular cancer come from the differences in male reproductive pathology between men from Denmark and those from Finland. Not only do Danish men have the lowest sperm counts in Europe, but they also exhibit high incidences of testicular cancer and malformations of the genital tract such as hypospadias. By contrast, the incidence of testicular cancer in Finland is

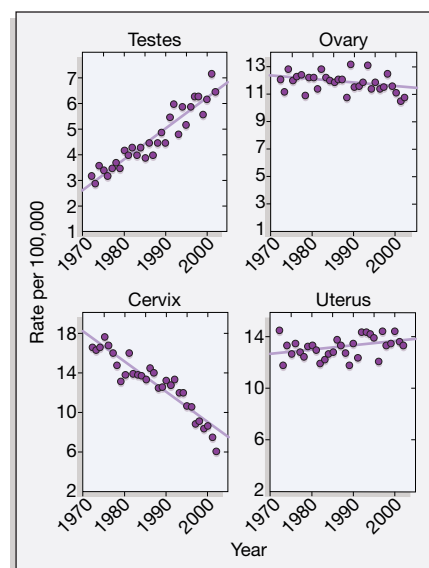


Figure 2 Trends in reproductive-tract cancers. As exemplified by these data from the New South Wales Central Cancer Registry³¹, during the past 30 years the incidence of testicular cancer has continued to rise whereas incidence of cancers of the female reproductive tract has remained constant (ovary, uterus) or has decreased (cervix). Similar trends have been seen in all developed countries where data are available.

nearly three times lower, genital malformations are rare and mean sperm counts are among the highest in the world¹⁸. We have no idea why the reproductive fate of men in these two countries is so different. One recent publication, however, has highlighted the powerful correlation between the incidence of maternal smoking during pregnancy and the relative incidence of testicular cancer across four Nordic countries (Sweden, Denmark, Norway and Finland)¹⁹.

“Environmental factors, whatever the route of exposure, can clearly affect the development and function of the male reproductive tract.”

Another facet of investigations into male reproduction is the possibility that damage to a father's sperm — either genetic, affecting the DNA sequence itself, or otherwise perturbing DNA function through so-called epigenetic mechanisms — can be responsible for diseases in his offspring, as well as being itself a cause of infertility or early loss of pregnancy. A frequently quoted example is the link between heavy paternal smoking and increased rates of childhood cancer⁶, thought to be mediated by oxidative damage to the DNA in the father's sperm. Oxidative injury results from an over-exposure to 'reactive oxygen species' — excited oxygen-containing molecules, generated as a by-product of cell metabolism and the intracellular processing of xenobiotics, that can attack and damage DNA.

With the traditional emphasis on the impact of cigarette smoke on somatic tissues (for instance in heart disease and lung cancer), we tend to forget that smoking can

also induce oxidative damage to the DNA in sperm, and thereby affect the health and wellbeing of the ensuing children. Moreover, because such damage is in the germ line, it can be transmitted to future generations. We are probably all carrying around in our genes the consequences of our great grandfather's pipe-smoking habit.

DNA damage in human sperm has also been associated with a reduction in overall pregnancy rates following natural conception. Moreover, such damage has been linked with impaired fertilization, disrupted development of the early embryo, and loss of pregnancy in assisted reproduction programmes⁶.

Xenobiotics and reproduction

What can be causing all of these male reproductive problems? Long-term impacts on human fertility are associated with our evolutionary heritage, in that poor sperm morphology is a burden we share with some of our close primate relatives, such as the gorilla. Also, there is a lack of selection for 'high-fertility' genes in countries that have gone through the demographic transition — that is, the transition from high birth and death rates, to low birth and death rates. The increasing availability of assisted conception clinics will further dilute selection for high-fertility genes and, in this context, the slow drift towards increased infertility in developed countries is inexorable.

But these gradual trends are being accelerated by environmental factors that seem to be having a particular impact on the male germ line: Fig. 3 provides examples of some of the main groups of xenobiotics. One of the most intensively researched groups is the environmental oestrogens: these phenolic compounds, found in plants but also in man-made products, competitively interact with the body's receptors for the natural oestrogen, a steroid hormone. Oestrogen is generally thought of as a female hormone. But normal male development and function also depend on it. One explanation for reduced sperm counts involves the capacity of environmental oestrogens to suppress production of a hormone (follicle-stimulating hormone) by the fetal pituitary gland. As this hormone stimulates the growth of Sertoli cells in the developing testes, the number of these cells is consequently decreased¹⁷. Sertoli cells rarely, if ever, replicate, and each cell can only support the

Compound	Source	Biological action
1,2-dibromo-3-chloropropane	Used as a pesticide, nematocide and soil fumigant. Exposure by ingestion of contaminated water or food, or breathing air near contaminated sites. Occupational exposure in the chemicals industry.	Disrupts sperm production and causes male sterility. Some evidence that paternal exposure increases the incidence of abortion but results are inconsistent across studies. In animal studies, mating to exposed males increases subsequent pregnancy loss.
Nonylphenol	Industrial surfactant and antioxidant released by degradation of ethoxylated nonylphenol derivatives. Common contaminant of aquatic environments.	Environmental oestrogen responsible for feminization of aquatic species, including fish and oysters. Both fetal and adult exposure causes testicular damage in rats. Can induce DNA damage in human sperm.
Genistein	Plant oestrogen, present in soybean products. High daily exposure (1–30 mg per day) particularly in Asian populations.	No convincing evidence that developmental exposure impairs male reproduction. Acute exposure does not affect human semen quality. But this and related plant oestrogens do induce DNA damage in human sperm.
Polycyclic aromatic hydrocarbons, including benzo[a]pyrene	Constituents of cigarette smoke.	Smoking induces modest reductions in semen quality. Heavy smoking results in oxidative base damage in sperm and can affect offspring (for instance by inducing childhood cancer). Animal studies indicate effects on sperm maturation and ability of sperm to establish a viable pregnancy.
Acrylamide	Intermediate in the synthesis of polyacrylamide, with widespread use in water treatment, paper making, ore processing and the manufacture of diverse products. Contaminant of fried food.	Disrupts male reproduction by many mechanisms, including impairment of sperm development and motility. Male exposure leads to pregnancy loss in mated females, and increased incidence of developmental defects and hereditary chromosome disorders.
Dioxins, for example 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD)	By-products of combustion and various industrial processes. Dietary exposure particularly through foods containing animal fats. Minor exposure from absorption or ingestion of contaminated dust.	Hormone disruptor that impairs sperm development and sexual maturation. Increased risk of testicular cancer and birth defects, but data are inconsistent. Induction of oxidative stress in sperm.

Figure 3 Xenobiotics under suspicion. Examples of chemicals implicated in adversely affecting the male reproductive system¹.

DNA or generating reactive oxygen species. The latter are created by a ping-pong type of activity termed 'redox cycling', whereby a given compound alternates between oxidized and reduced states. During this process, electrons are transferred to oxygen to produce the superoxide anion. The latter can then create a state of oxidative stress through a complex series of secondary reactions that result in damage to both the sperm and its DNA⁶. Exposure of human sperm to oestrogenic compounds such as diethylstilbestrol, phyto-oestrogens (equol, genistein and daidzein), industrial surfactants (nonylphenol) and natural oestrogens (oestradiol-17 β) can induce significant DNA damage through mechanisms that seem to involve oxidative stress²¹. Most DNA damage in the sperm of infertile males also seems to have been caused by oxidative damage, resulting in high levels of the oxidized DNA product, 8-hydroxy-2'-deoxyguanosine²².

At present, we know very little about the nature of the xenobiotic-metabolizing enzymes in the male germ line, and thus the potential that different groups of compounds have for inducing genetic damage by oxidative, or other, mechanisms is uncertain. Experimentally, we know that a state of oxidative stress can be induced in the testes by exposure to common xenobiotics such as nonylphenol or dioxin²³. But the biochemical mechanisms underpinning this activity remain unclear.

Epidemiological clues

The epidemiology literature offers some clues about the germ-cell toxins that can damage the offspring of affected males. As well as the link between paternal cigarette-smoking and childhood cancer, the chance of contracting testicular cancer can depend on paternal occupation. If a man works in the wood-processing industry, then the risk of his son developing testicular cancer is more than ten times that for the average male; for the sons of men working in the metal industry, the risk is about six times the average²⁴. There are also data linking childhood cancer with paternal exposure to hydrocarbons in various forms (benzene, paint, methyl ethyl ketone, plastic and resin fumes, and different types of solvents)²⁵. In addition, an increased risk of contracting leukaemia has been reported in children whose fathers are automobile, truck or aircraft mechanics.

differentiation of a finite number of sperm. So a reduction in the size of this cell population should have an irreversible impact on male germ-cell development.

An alternative possibility is that environmental oestrogens impair Leydig cell development or function, thereby affecting the generation of testosterone by the testes. These environmental oestrogens might also inhibit the action of the molecular receptors for testosterone and other male hormones, or suppress the expression of growth factors (such as insulin-like growth factor-3) within the fetal testes¹⁷. The long list of weakly oestrogenic factors that can act as hormone disruptors includes nonylphenol, plant-derived oestrogens such as genistein, and dioxins (Fig. 3), as well as DDT, furans and the insecticides dieldrin and aldrin. Some of the more toxic compounds (DDT, dieldrin and aldrin) have been banned from most industrialized countries since the 1970s. But the bans are not universal, and these compounds continue to accumulate in the global environment through food imports and contaminated air or water.

The hormone-disruptor hypothesis is certainly plausible, but there are several difficulties with the argument that have yet to be

addressed adequately. First, environmental oestrogens exhibit only weak biological activity; DDT for example has a bioactivity level that is 100,000 to 1,000,000 times less than the human oestrogen oestradiol-17 β . Potency calculations suggest that the daily birth-control pill involves exposure to about

ten billion times more oestrogenic activity than the dietary intake of organochlorine pesticides, such as DDT or dioxin-like compounds²⁰. In addition, the male fetus is exposed to very high levels of potent placental oestrogens during the course of normal pregnancy and yet

develops normally. Finally, the male offspring of women exposed during pregnancy to a powerful synthetic oestrogen, diethylstilbestrol, failed to exhibit a statistically significant change in the incidence of testicular cancer or infertility. The observed effects (small testes, cryptorchidism, epididymal cysts) are much less frequent than would have been expected if the oestrogenicity of the compound were the only determinant of its developmental toxicity. Other mechanisms must be involved.

One possibility is that some of these environmental oestrogens can be metabolized further to molecules (quinones) that can cause cellular damage by either binding to

"We tend to forget that smoking can induce damage to the DNA in sperm, and thereby affect the health and wellbeing of ensuing children."

These associations suggest that there is a chain that links paternal exposure to xenobiotics with genetic or epigenetic DNA damage to the father's sperm, and with adverse consequences for his offspring. Although the epidemiology literature is not always consistent (Box 1), these links have been observed in animal models where paternal exposure to a wide variety of potential toxicants (including acrylamide, cyclophosphamide and urethane) is significantly associated with increased incidences of spontaneous abortion and birth defects in the offspring²⁶.

Clinical studies in this area would be much improved if we had a better understanding of the way in which xenobiotics are metabolized in the male germ line and the kinds of DNA damage induced. Metabolism of organic xenobiotics involves the initial biochemical modification of a given compound (phase 1) followed by a conjugation reaction that links the modified compound to a carrier molecule in preparation for excretion (phase 2). The ability of individual males to metabolize and link xenobiotics in this manner depends heavily on genetically determined variations in the enzymes responsible for the biotransformation reactions. Knowing more about those enzymes, and the relative sensitivities of subjects with specific genotypic profiles, will help epidemiologists untangle the relationships between toxicant exposure and adverse reproductive outcomes. This could be the dawning of the age of 'reproductive pharmacogenomics', in which a male's reproductive susceptibility to a xenobiotic can be predicted from his profile for key enzymes such as cytochrome P450 (phase 1) and glutathione-S-transferase (phase 2). There is already evidence to suggest that variation in cytochrome P450 enzymes affects male fertility²⁷.

Age and mobile phones

Xenobiotics are not the only factors that can induce oxidative DNA damage in the male germ line. Age is another: as a man ages, his sperm count may not change significantly but the amount of DNA damage in his sperm increases dramatically: the amount of DNA damage in sperm of men aged 36–57 is three times that of men below the age of 35 (ref. 28). This age-dependent increase in

DNA damage may contribute to the incidence of childhood diseases that increase with paternal age, including complex conditions such as schizophrenia, and genetic disorders such as the achondroplasia caused by defective bone growth²⁹.

Although the risk to the individual is low, such associations could become more significant if assisted-conception protocols that do not exclude DNA-damaged sperm, such as ICSI (intracytoplasmic sperm injection), continue to be used to address age-related declines in human fertility—in older men starting second families, for example.

Radiofrequency electromagnetic radiation may be another cause of damage to the male germ line, given preliminary reports of DNA damage in the sperm of mice exposed to mobile-phone radiation³⁰. Such results are bound to generate widespread publicity, but the data are still much too limited to draw any conclusions.

The future

So we are faced with a situation where semen quality is apparently declining and pathologies of the male reproductive tract are rising; moreover, 3–6% of the population in most developed countries is now produced by assisted conception. To some extent the slide towards lower fertility is a consequence of lifestyle choices (more young adults deliberately delaying parenthood or choosing a child-free future) and a lack of selection pressure on high-fertility genes, which we are powerless to prevent. But some of the reproductive pathologies we are seeing are a biological response to factors in the environment that are affecting every aspect of human reproduction from fertilization of the egg, through fetal development of the reproductive system, to child health.

The developmental consequences of environmentally mediated DNA damage to sperm include impaired embryonic development, abortion and the induction of abnormalities in the offspring such as childhood or testicular cancer. But although we know that environmental factors can induce severe damage in male germ cells, we know little more than that. The questions of the kinds of molecular structure that induce

such damage, the nature of the damage induced, and the mechanisms by which such damage affects embryonic development all require urgent attention. ■

R. John Aitken is in the Discipline of Biological Sciences, School of Environmental and Life Sciences, University of Newcastle, Callaghan, New South Wales 2308, Australia, and Peter Koopman is in the Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia. Both are members of the ARC Centre of Excellence in Biotechnology and Development, University of Newcastle. Sheena E. M. Lewis is in Obstetrics and Gynaecology, Queen's University, Belfast BT12 6BJ, UK.

e-mail: jaitken@mail.newcastle.edu.au

1. Robaire, B. & Hales, B. F. *Advances in Male Mediated Developmental Toxicity* (Kluwer/Plenum, New York, 2003).
2. Klaassen, C. D. Casarett & Doull's *Toxicology: The Basic Science of Poisons*, 6th edn (McGraw-Hill, New York, 2001).
3. Whorton, D., Milby, T. H., Krauss, R. M. & Stubbs, H. A. *J. Occup. Med.* **21**, 161–166 (1979).
4. Oliva, A., Spira, A. & Multigner, L. *Hum. Reprod.* **16**, 1768–1776 (2001).
5. Hill, K. A. et al. *Environ. Mol. Mutagen.* **43**, 110–120 (2004).
6. Aitken, R. J. *Reprod. Fertil. Dev.* (in the press).
7. Skakkebaek, N. E. *Int. J. Androl.* **27**, 189–191 (2004).
8. Fisher, J. S., Macpherson, S., Marchetti, N. & Sharpe, R. M. *Hum. Reprod.* **18**, 1383–1394 (2003).
9. Guillelte, L. J. et al. *Environ. Health Perspect.* **102**, 680–688 (1994).
10. Semenza, J. C. et al. *Environ. Health Perspect.* **105**, 1030–1032 (1997).
11. Jobling, S. et al. *Aquat. Toxicol.* **66**, 207–222 (2004).
12. Carlsen, E., Giwercman, A., Keiding, N. & Skakkebaek, N. E. *Br. Med. J.* **305**, 609–613 (1992).
13. Swan, S. H., Elkin, E. P. & Fenster, L. *Environ. Health Perspect.* **108**, 961–966 (2000).
14. Irvine, S., Cawood, E., Richardson, D., MacDonald, E. & Aitken, J. *Br. Med. J.* **312**, 467–471 (1996).
15. Andersen, A. G. et al. *Hum. Reprod.* **15**, 366–372 (2000).
16. Handelsman, D. J. *Reprod. Fertil. Dev.* **13**, 317–324 (2001).
17. Sharpe, R. M. *Int. J. Androl.* **26**, 2–15 (2003).
18. Slama, R. et al. *Hum. Reprod.* **17**, 503–515 (2002).
19. Pettersson, A. et al. *Int. J. Cancer* **109**, 941–944 (2004).
20. Safe, S. H. *Environ. Health Perspect.* **103**, 346–351 (1995).
21. Anderson, D. A. et al. *Mutat. Res.* **544**, 173–187 (2003).
22. Kodama, H., Yamaguchi, R., Fukuda, J., Kasai, H. & Tanaka, T. *Fertil. Steril.* **68**, 519–524 (1997).
23. Chitra, K. C. & Mathur, P. P. *Indian J. Exp. Biol.* **42**, 220–223 (2004).
24. Knight, J. A. & Marrett, L. D. *J. Occup. Environ. Med.* **39**, 333–338 (1997).
25. Shu, X. O. et al. *Cancer Epidemiol. Biomarkers Prev.* **8**, 783–791 (1999).
26. Anderson, D. *Adv. Exp. Med. Biol.* **518**, 11–24 (2003).
27. Schuppe, H. C. et al. *Andrologia* **32**, 255–262 (2000).
28. Singh, N. P., Muller, C. H. & Berger, R. E. *Fertil. Steril.* **80**, 1420–1430 (2003).
29. Kühnert, B. & Nieschlag, E. *Hum. Reprod. Update* **10**, 327–339 (2004).
30. Aitken R. J., Bennetts, L. E., Sawyer, D., Wiklendt, A. M. & King, B. V. *Int. J. Androl.* (in the press).
31. Tracey, E. A., Chen, W. & Sitas, F. *Cancer in New South Wales: Incidence and Mortality, 2002* (NSW Cancer Council, Sydney, 2004).

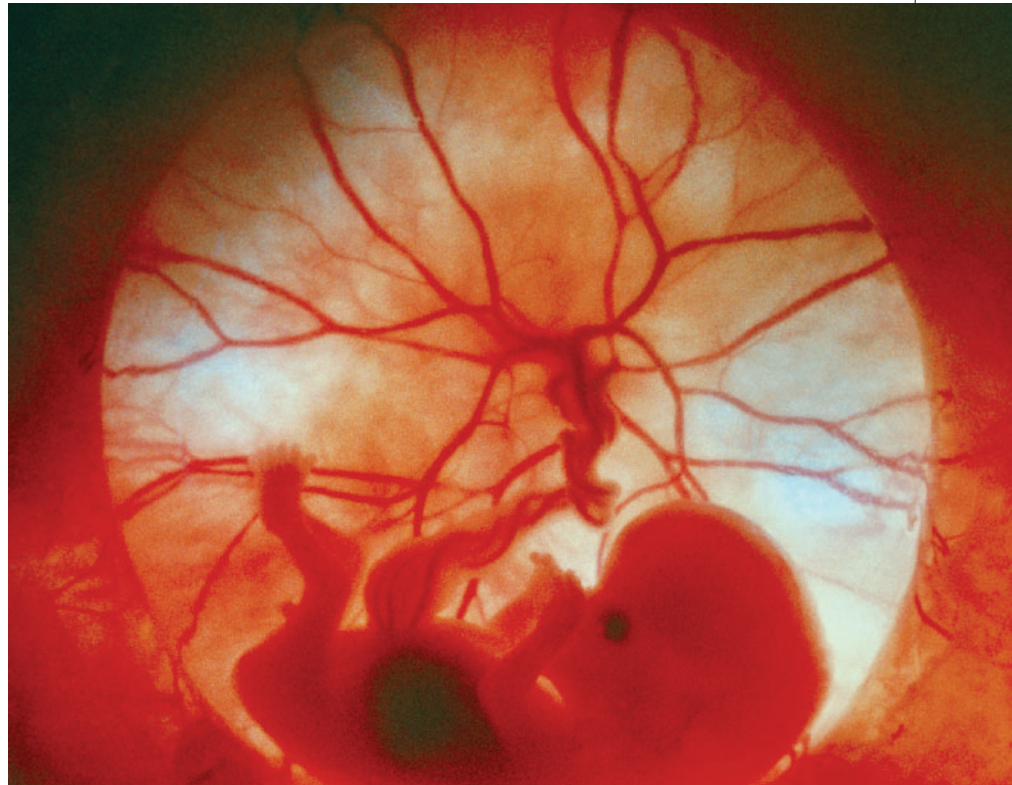
Resourceful imprinting

A child's genes are not all equal: in some cases, the copy from either the mother or the father is turned off. This affects the child's ability to acquire resources in the womb, after birth, and perhaps throughout life.

**Miguel Constância, Gavin Kelsey
and Wolf Reik**

Children spend about a third of their lives acquiring resources from their parents. Parental investment often continues after children finish full-time education and establish their own families. Resources at this stage vary from money to advice; earlier in life they include the teaching of complex social behaviours, social, intellectual and language skills, and knowledge about our world. Although these are not acquired with the mother's milk, the rapid development of the necessary motor and sensory skills does depend on feeding and milk metabolism during the earliest part of post-natal life. How well the newborn baby fares depends in turn on how well it was nourished in the womb, allowing it to grow and develop the organs it needs for life after birth.

This matters because the more resources parents invest in their offspring, the better the offspring's chances of surviving and being attractive as a sexual partner, and therefore of reproducing. Contrast this enormously drawn-out process, particularly in primates, with species of fish, for example, that leave their masses of young to fend for themselves, with the result that most end up as food for the local community instead of reproducing. So, in higher mammals in particular, genes in offspring that make them acquire and store resources, and genes in parents that make them donate resources to their offspring, are crucial for a long and reproductively successful life.



But the tangible contributions that mothers and fathers make to their children are very different — indeed, some might ask what fathers do beyond donating sperm. During a child's development in the womb, and post-natal feeding up to weaning, the father is clearly not a major direct player. Because of this fundamental asymmetry, resource-acquisition genes in offspring have different, and conflicting, strategies depending on which parent they come from. (It is important to distinguish genes in the parents, that is, maternal or paternal genes, from maternally or paternally inherited genes in the offspring.)

These genetic asymmetries are the subject of kinship theory, also known as genetic conflict theory¹. This proposes that paternally inherited resource-acquisition genes are 'interested' in extracting as many resources from the mother as possible — in the form of nutrients through the placenta during gestation, or milk after birth, for example — as this normally costs the father little. The possibility that the same mother will bear offspring by different males strengthens this drive.

Maternally inherited genes, by contrast, are not interested in exhausting maternal resources prematurely, but rather in ensuring that these resources are dished out even-handedly to all offspring.

Imprinting in theory

How can maternally and paternally inherited genes be made to behave differently? The only real option is to switch one gene copy off. This is achieved by an epigenetic mechanism of gene regulation — one in which the gene's sequence is unchanged, but its labelling with certain chemical groups is altered. The mechanism in question is termed genomic imprinting^{2,3}, and it ensures that the relevant parental genes are marked in the germ cells (egg and sperm) to be active or inactive in the offspring (Box 1, overleaf). The conflict between parental gene sets works at different levels. For instance, a gene that confers a growth advantage when inherited from the father might have been subject during evolution to silencing of the maternal copy. Conflict

EDELMANN/SPL

Box 1 Control of imprinting

The distinguishing feature of imprinted genes is that the two parental gene copies, or alleles, are expressed differently. In many cases, one allele is silent in most tissues throughout development. This means that the alleles retain a memory of their parental origin. In molecular terms, imprinted genes acquire different marks in the sperm and egg, and these marks must be heritable through subsequent cell divisions^{2,3}. Part of the marking is DNA methylation — the covalent addition of methyl groups to cytosine residues in CpG dinucleotides, which occurs in imprinting control regions (ICRs) of DNA.

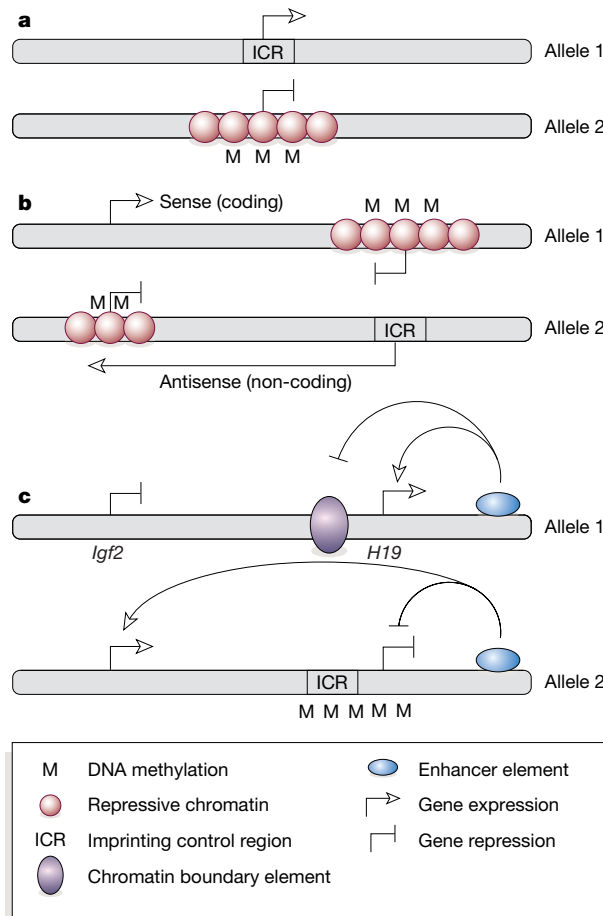
DNA methylation on imprinted genes is established by the *de novo* methyltransferase enzyme Dnmt3a, together with the Dnmt3l protein²⁹. After fertilization, the maintenance methyltransferase Dnmt1 ensures the faithful copying of methylation marks through subsequent cell divisions. DNA methylation marks are usually erased only when the alleles encounter the germline again, in the primordial germ cells, before they are re-established according to the sex of the individual.

DNA methylation and associated differences in chromatin — the way

in which DNA is packaged — are read after fertilization to ensure that the correct allele is expressed. This occurs in different ways depending on the gene^{2,3} (see figure). In the simplest case (a), DNA methylation and repressive chromatin may directly overlie one allele's promoter region (from which gene expression begins), silencing that allele (allele 2 in this case).

In an indirect mechanism (b), the ICR contains the promoter of a non-protein-coding gene, which can be expressed to produce an 'antisense' RNA; this RNA then somehow (perhaps through methylation) silences the protein-coding imprinted gene on the same chromosome. On the other chromosome, methylation of the ICR ensures that the antisense RNA is not expressed.

In a third mechanism (c), the ability of shared activating sequences (enhancers) to activate one or other imprinted gene is determined by a chromatin boundary element present on the unmethylated allele between the two genes, as exemplified by *Igf2* and *H19*. Finally, reading mechanisms may vary in different tissues, resulting in tissue-specific imprinting of some genes.



may also operate at the protein level: thus, the insulin-like growth factor type 2 receptor (Igf2r), which is expressed only from the maternal gene copy, encodes a receptor protein that degrades excess Igf2 — a paternally expressed growth enhancer.

Although only some 80 of the 30,000 genes in the human genome are currently known to be imprinted^{4,5}, they undoubtedly have key roles in resource acquisition. Crucially, paternally and maternally expressed imprinted genes also affect resource transfer in the expected direction — genes expressed from the paternally inherited copy generally increase resource transfer to the child, whereas maternally expressed genes reduce it⁶.

If imprinted genes are indeed intimately linked with acquiring resources from parents, we would not expect any in the fish that leave their offspring to their own devices after conception. Indeed, early indications provide little evidence of genomic imprinting in fish, amphibians, reptiles and birds⁷. But imprinting does also exist in seed plants, the endosperm tissue of which is the equivalent of a placenta to feed the embryo. And, although imprinted genes concerned with fetal resources seem particularly prominent in

placental mammals (which are divided into marsupials, and others, including humans, that are collectively known as eutherians), genes involved in suckling could also be imprinted in egg-laying mammals (monotremes). Imprinted genes affecting behaviour might also operate in many other situations in which there is parental asymmetry in investment, especially in higher primates.

If imprinted genes are crucial in mammalian development, one would also expect mutations in these genes to cause diseases, and that these diseases might be recognizable by their exclusive maternal or paternal transmission in families. There is indeed a growing number of known imprinting disorders⁵, many of which affect fetal growth (resulting in babies that are too small or too large), hormone systems after birth, or adult behaviour (Fig. 1). As imprinting is an epigenetic mechanism of gene regulation, epimutations — mistakes in maintaining epigenetic marks — also cause imprinting disorders.

Imprinting in the womb

Kinship theory predicts that selective forces will favour parent-specific gene expression most strongly during mother–offspring

interactions in the womb and soon after birth¹. But how does imprinting affect life *in utero*?

The human fetus is tucked neatly away in the womb for nine months, feeding off the mother's nutrient supplies. This is possible because of the evolution of an organ for exchanging nutrients, gases and waste products between mother and fetus — the placenta. The human placenta is remarkable, providing 11 square metres of exchange surface at the time of birth, and is almost as efficient as the lungs in gas exchange. What are the advantages of having acquired such an organ? For a start, the placenta allows for a prolonged gestation period, which in turn allows larger, more-developed young at birth, with more advanced central nervous systems. Like eutherian mammals, marsupials also benefit from a period of life in the womb supported by a placenta, although the period is briefer, the placenta is much more rudimentary, and the baby is less developed at birth.

A fetus, although entirely dependent on its mother's nutrients, is not just a passive recipient, but influences its own development and growth. Indeed, it subverts many

Disorder	Effect	Imprinted genes suspected or known to be affected	Expressed gene copy
Intra-uterine growth			
Beckwith–Wiedemann syndrome	Fetal and postnatal overgrowth; excessively large organs; predisposition to tumours	<i>IGF2</i> (encoding a growth factor) <i>CDKN1C</i> (encoding a cell-division regulator)	Paternal Maternal
Silver–Russell syndrome	Severe intra-uterine growth restriction	Maternal uniparental disomy and duplications of chromosome 7	
Pre-eclampsia	Pregnancy-associated hypertension, often accompanied by intra-uterine growth restriction	Linkage studies suggest involvement of maternally expressed imprinted genes in some families	
Behaviour and brain			
Prader–Willi syndrome	Moderate mental retardation; severe obesity; short stature; poor muscle tone	Numerous imprinted genes on chromosome 15	Paternal
Angelman syndrome	Severe motor and mental retardation; paroxysms of laughter; autistic-like behaviour	<i>UBE3A</i> (encoding a protein-degradation regulator)	Maternal
Turner syndrome (monosomy X)	Affects females only; associated with a characteristic neurocognitive profile, short stature and ovarian failure	Enhanced social cognitive skills in patients inheriting the paternal, rather than maternal, X chromosome may indicate imprinting	
Schizophrenia	Perceived distortions of reality; disturbance of thought and language; withdrawal from social contact	Some forms of schizophrenia show lower age of onset after paternal inheritance	
Maternal behaviour defects (in mice)	Lack of maternal postnatal care of offspring	<i>Peg3</i> (encodes a DNA-binding protein) <i>Peg1</i> (encodes an enzyme of the α/β -hydrolase family)	Paternal Paternal
Hormones and metabolism			
Albright hereditary osteodystrophy	Short stature; round face; obesity; mental retardation; subcutaneous calcification	<i>GNAS</i> (encodes a G-protein subunit)	Maternal (tissue specific)
Pseudohypoparathyroidism 1A	As above, accompanied by resistance to parathyroid hormone and other hormones	Occurs only on maternal transmission of inactivating <i>GNAS</i> mutations	Maternal (tissue specific)
Transient neonatal diabetes mellitus	Pancreatic insufficiency and low secretion of insulin during fetal life; intra-uterine growth restriction	<i>PLAGL1</i> (encodes a DNA-binding protein)	Paternal

Figure 1 Imprinting and disease. Several disorders (not all of which are shown) are caused by the absence or misexpression of imprinted genes. The mechanisms responsible may be genetic, such as gene deletions, or the inheritance of both chromosome sets — or parts of them — from just one parent instead of two (uniparental disomy). The mechanisms might also be epigenetic, involving for instance alterations in imprint marks or how they are ‘read’ (see Box 1). Most imprinting-associated defects arise from the loss of function of imprinted genes. A few result from increased

‘dosage’ of an imprinted-gene product, because of uniparental disomy or re-activation of the normally silent allele (loss of imprinting). Many other diseases may involve imprinted genes — they might, for instance, contribute to cancer — but this is unproven. The well-defined imprinting disorders are quite rare, affecting fewer than one person in 10,000. See ref. 30 for a detailed description of these diseases and their genetic causes; see also refs 4, 5, 8, 22, 31. Human genes are in upper-case, mouse genes in lower-case.

of the mother’s physiological activities to its own end, to ensure adequate mobilization of nutrients and oxygen. So it is unsurprising that the asymmetry of interests of paternally and maternally inherited genes — the desire to make a large, fit baby versus the desire to withhold resources for future offspring — is most tangible during life in the womb^{1,8}. One area for competition is in the placenta, over the control of the supply of nutrients.

The placenta comprises two components: a fetal portion that develops from trophoblast derivatives (the first lineage specified in the early embryo) and a maternal portion derived from the inner layer of the uterine wall. This juxtaposes fetal and maternal blood vessels for efficient nutrient transfer. The role of imprinting specifically in the placenta has been difficult to test by the traditional means of knocking out genes in experimental organisms and studying the effects, mainly because most imprinted genes are also expressed in the fetus, where

they may have additional roles (see below). Nonetheless, knock-outs of several imprinted genes in mice do affect the growth, thickness and organization of the placental tissues interposed between mother and fetus⁶ (Fig. 2, overleaf).

The direction of the effects supports the notion of competition — eliminating genes that are expressed from the paternal copy decreases the surface area for the exchange of nutrients; knocking out genes expressed from the maternal copy increases it. For example, removal of *Igf2* (the product of a paternally expressed gene) from the cells of the exchange barrier is enough to cause thickening of the barrier and decreased nutrient exchange⁹. A future challenge is to understand the impact of imprinting in other areas that may influence nutrient supply, such as fetal and maternal blood flow and the transport of nutrients themselves. So far, there are hints that imprinted genes might be involved in the development of the placental

blood vessels, and there are several imprinted genes that code for nutrient transporters (Fig. 2).

Another arena for the different interests of parental genomes to play out is the fetus, at the level of the demand for resources. Growth and cell proliferation impose demands on maternal provision. Thus, paternally inherited, demand-enhancing genes make babies grow, and increased fetal demand can be signalled back to the placenta, which can respond with increased supply¹⁰. Understanding such signalling systems between the fetus, the placenta and the mother will be important, particularly as there are situations in humans in which growth of the baby is compromised (Fig. 1; see below).

The connection between mother and fetus via the placenta, and then through suckling, is unique to mammalian reproduction. As imprinted genes have so far been identified only in placental mammalian species among vertebrates, it is possible that imprint-

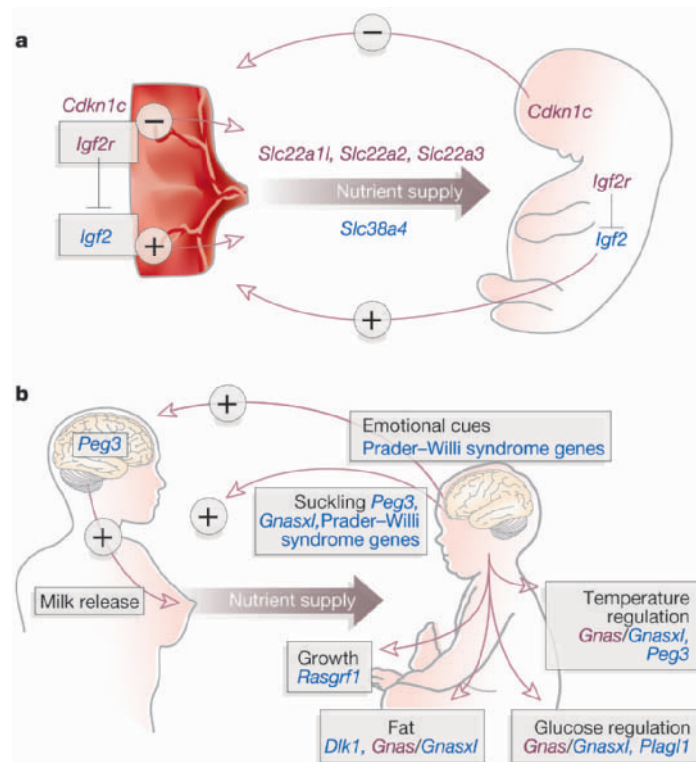


Figure 2 Effects of imprinted genes on resource acquisition by offspring⁶. Imprinted genes that are expressed from the maternally derived copy are in purple; those expressed from the paternally derived copy are in blue. a, Growth of the fetus — promoted by, for example, *Igf2*, a paternally expressed imprinted gene — may signal increased demand to the placenta. *Igf2* also increases the placenta's nutrient-transport capacity^{9,10}. Several nutrient-transporter-encoding genes (*Slc* genes) are also imprinted. Maternally expressed genes (such as *Igf2r* and *Cdkn1c*) may reduce nutrient supply or demand⁶. b, Imprinted genes might also control resource provision after birth, acting in the mother's brain to regulate milk release, or in the infant to regulate nipple attachment, suckling and feeding behaviours, including emotional interactions with the mother. Within the infant, imprinted genes may also affect the allocation of acquired resources into growth, fat reserves and homeostatic mechanisms, such as glucose and temperature regulation. Most genes shown were discovered through experiments in mice, but might be common to humans.

ing coevolved with placentation, or itself drove the evolution of the placenta. A fundamental connection between imprinting and placentation is supported by the finding that erasing all genomic imprints results in catastrophic development of extra-embryonic tissues and the placenta in mice^{11,12}. And in humans, a failure to set imprints in the female germline gives rise to a hydatidiform mole — a tumour-like mass resulting from uncontrolled growth of placental material¹³. So far, however, only a single imprinted gene, coding for the gene-transcription factor *ASCL2*, has been shown to be essential for forming a placenta, in mice¹⁴.

Can we make predictions about the strength of imprinting in different mammals? We might expect imprinting in species with a simple placenta, such as marsupials, to be weak, and not as widespread as in species with a more complex, longer-lived and functionally more important placenta. It remains to be seen whether this is true.

Imprinting after birth

Does imprinting matter after birth to the extent that it does *in utero*? There is every reason to believe that it does. For instance, one of the first human genetic diseases found to be caused by mutations in imprinted genes was Prader-Willi syndrome (PWS; Fig. 1). This is a complex childhood disorder that affects the nervous and hormonal systems and ultimately leads to excessive weight gain. It arises from the loss of activity of several imprinted genes, found on chromosome 15, that are usually expressed from the paternally inherited copy. Another reason for sus-

pecting that imprinting is still important after birth is that the epigenetic marks imprinted on genes in the male and female germ cells (Box 1) persist into adulthood, and many imprinted genes continue to be expressed from a single parental copy. Epimutations sustained during development may thus deregulate imprinted genes throughout life, as there seems to be no way to correct such changes — at least until the gene passes through the germline once again.

Birth is a traumatic process for mother and baby alike, requiring a host of physiological adaptations to meet the challenges of the postnatal environment and the need of the infant to take more responsibility for its own well-being.

After birth, we must find our own sources of nutrition, first from the mother's breast, later from supplemental or foraged foods (Fig. 2). We must learn to recognize when we are hungry and when sated. For the first time we need to regulate numerous metabolic processes, such as ensuring that blood sugar levels remain adequate to nourish the brain. Our bodies must make decisions about how much energy to devote to the demanding processes of growth and development, and how much to other essential functions, such as keeping ourselves warm. On top of this, we are subject to far greater influences from the environment and from competition with siblings (and possibly half-siblings).

This complexity makes it more difficult to predict what imprinting is doing after birth. But we might expect imprinted genes

to be most influential in situations in which investment in young comes at a direct cost to mothers and their ability to invest in current or future offspring, and where there may be competition within or between the offspring of unrelated fathers.

The obvious place to look first is suckling, while the infant is still nutritionally dependent on, and demands a great deal from, mum. Returning to our example above, PWS children are born profoundly listless, show little interest in feeding and suckle poorly for the first months of their lives; their voracious appetites and consequent obesity develop from weaning onwards. Moreover, the expression pattern of one imprinted gene with an apparently key role in neonatal adaptations — coding for the paternally expressed *XLas* protein — coincides quite remarkably with the centres in the brain that innervate the apparatus involved in suckling¹⁵.

It is unlikely that maternally expressed imprinted genes exist to prevent suckling, but the antagonism among imprinted genes may continue after birth. A striking example comes from mice in which the maternally expressed *Gnas* gene is knocked out. One of the processes *Gnas* is involved in is metabolism. It codes for a subunit of a G protein — a common type of molecular switch — that transduces signals from hormone receptors to an enzyme known as adenylate cyclase, which in turn promotes the production of the molecule cyclic AMP (cAMP). One effect is to stimulate heat production in fat tissue. If

"It is unsurprising that the asymmetry of interests of paternally and maternally inherited genes is most tangible during life in the womb."

mice receive a mutant copy of this gene from mothers, they develop a depressed metabolic rate, increase their storage of lipid in fat tissues, and become relatively obese. Surprisingly, mice inheriting the mutation in the paternal copy have an elevated metabolic rate, reduce stored lipid and remain lean¹⁶.

These effects seem paradoxical, until we learn that the mutation also hits a paternally expressed gene in this complicated imprinted genetic region; this gene encodes a variant G-protein subunit that may attenuate cAMP signalling¹⁵. Why should imprinting have evolved to control metabolic rate? Maintaining body temperature in young mammals is costly, and it may be that selfish paternal genes have sought to invest less in heat production (and so more in growth) by free-riding on the heat production of littermates¹⁷.

Intriguingly, imprinted genes in the mothers themselves can intervene in this conflict. The *Peg3* gene, the paternally derived copy of which is expressed, dictates not only how well pups thrive, controlling their suckling and weight gain, but also how good mothers are at provisioning their offspring¹⁸. Female mice lacking *Peg3* put on fewer reserves during pregnancy and show poor maternal care and impaired milk release. Combining the *Peg3* deficit in both pups and mums is particularly devastating, raising the intriguing prospect of co-adaptation of this gene in its effects in mother and infant.

The extent to which imprinting controls postnatal physiology and metabolism in mammals remains to be fully explored, and it may be premature to expect a unifying explanation of its effects after birth. That its influence may be quite wide-ranging is suggested by the imprinting of such a central player as a G-protein subunit, and the resistance to numerous hormones caused by loss of the maternal copy of *GNAS* in humans (Fig. 1). Potent effects on the development of fat tissue and on pancreatic function are also associated, respectively, with the imprinted genes *Dlk1* (ref. 19) and *PLAGL1* (ref. 20), mutations in the latter being responsible for an inherited form of diabetes in humans. The imprinting of the gene encoding D3 — the enzyme that inactivates thyroid hormone — also points to a role for imprinting in controlling the activity of a potent developmental

and metabolic hormone, certainly *in utero*, but possibly also after birth²¹. Given the range of physiological effects, particularly those underlying common human diseases, this is an important area for future investigation.

Perspectives

Imprinting is a genetic mechanism that regulates the demand, provision and use of resources in mammals. Its influence extends from the fetus (particularly in eutherian mammals), through the suckling period (in all mammals, perhaps including monotremes), to after weaning. At this stage, complex behaviours and social cognitive processes may be affected, particularly in higher primates²². For example, in both

humans and mice there seem to be imprinted genes on the X chromosome that regulate social or cognitive behaviours²³ (W. Davies *et al.*, personal communication). And children with Angelman's syndrome, which arises from the opposite

imprinting defects to those described in Prader–Willi syndrome, laugh and smile often and have a generally happy disposition — an indication, perhaps, that paternal genes may be exploiting emotional cues to increase the attention and resources the child receives from the mother²⁴. Moreover, the paternally expressed *Peg3* gene affects the behaviour of the mother as well as fetal growth¹⁸. Once we recognize this common theme of resource regulation, we may be able to make sense of imprinting-associated disorders that otherwise seem to defy explanation.

The fact that many imprinted genes are expressed in the placenta, the fetus and the brain — both neonatal and adult — suggests that imprinting might accompany our life from fetus to baby to adult, and from child to parent. That maternal (and perhaps paternal) genes, and maternally and paternally inherited genes in offspring, deploy different strategies in this game is both fascinating and scary — fascinating because it provides unique insights into mammalian biology and reproduction, and scary because the conflicting strategies can cause disease.

This is particularly important because imprinting can be altered by both classical mutations and epimutations, which may be more common. Epimutations contribute to human disease²⁵; how they arise is not well

understood, but one trigger is apparently the culture and manipulation of early embryos, as increased rates of imprinting disorders have been observed after assisted reproduction techniques, such as *in vitro* fertilization and intracytoplasmic sperm injection, in humans²⁶.

Changes in imprinting might also explain why restricted growth of the fetus in the womb (intrauterine growth restriction) can be associated with an increased risk of cardiovascular disease, diabetes and mental defects later in life (the 'fetal programming' hypothesis of adult diseases²⁷). It is not yet known how, but environmental influences during life in the womb might affect epigenetic marks, including perhaps those in imprinted genes, shifting resource regulation over the course of a lifetime²⁸.

Miguel Constância, Gavin Kelsey and Wolf Reik are in the Laboratory of Developmental Genetics and Imprinting, The Babraham Institute, Cambridge CB2 4AT, UK.

e-mails: miguel.constancia@bbsrc.ac.uk;

gavin.kelsey@bbsrc.ac.uk; wolf.reik@bbsrc.ac.uk

- Wilkins, J. F. & Haig, D. *Nature Rev. Genet.* **4**, 359–368 (2003).
- Reik, W. & Walter, J. *Nature Rev. Genet.* **2**, 21–32 (2001).
- Ferguson-Smith, A. C. & Surani, M. A. *Science* **293**, 1086–1089 (2001).
- Beechey, C. V., Cattanach, B. M. & Blake, A. Mouse Imprinting Data and References (www.mgu.har.mrc.ac.uk/imprinting/imprinting.html).
- Morison, I. M., Paton, C. J. & Cleverley, S. D. *Nucl. Acids Res.* **29**, 275–276 (2001); www.otago.ac.nz/LGC
- Reik, W. *et al. J. Physiol. (Lond.)* **547**, 35–44 (2003).
- Killian, J. K. *et al. Mol. Cell* **5**, 707–716 (2000).
- Tycko, B. & Morison, I. M. *J. Cell. Physiol.* **192**, 245–258 (2002).
- Sibley, C. *et al. Proc. Natl Acad. Sci. USA* **101**, 8204–8208 (2004).
- Constância, M. *et al. Nature* **417**, 945–948 (2002).
- Surani, M. A., Barton, S. C. & Norris, M. L. *Nature* **308**, 548–550 (1984).
- McGrath, J. & Solter, D. *Cell* **37**, 179–183 (1984).
- Judson, H. *et al. Nature* **416**, 539–542 (2002).
- Guillemot, F. *et al. Nature Genet.* **9**, 235–242 (1995).
- Plagge, A. *et al. Nature Genet.* **36**, 818–826 (2004).
- Yu, S. *et al. J. Clin. Invest.* **105**, 615–623 (2000).
- Haig, D. *Annu. Rev. Genet.* **38**, 553–585 (2004).
- Curley, J. P. *et al. Proc. R. Soc. Lond. B* **271**, 1303–1309 (2004).
- Moon, Y. S. *et al. Mol. Cell. Biol.* **22**, 5585–5592 (2002).
- Ma, D. *et al. J. Clin. Invest.* **114**, 339–348 (2004).
- Tsai, C. E. *et al. Curr. Biol.* **12**, 1221–1226 (2002).
- Isles, A. & Wilkinson, L. *Trends Cogn. Sci.* **4**, 309–318 (2000).
- Skuse, D. H. *et al. Nature* **387**, 705–708 (1997).
- Brown, W. M. & Consedine, N. S. *Med. Hypoth.* **63**, 377–385 (2004).
- Suter, C. M., Ward, R. L. & Martin, D. I. *Nature Genet.* **36**, 497–501 (2004).
- Maher, E. R., Afnan, M. & Barratt, C. L. *Hum. Reprod.* **18**, 2508–2511 (2003).
- Barker, D. J. *Trends Endocrinol. Metab.* **13**, 364–368 (2002).
- Waterland, R. A. & Jirtle, R. L. *Nutrition* **20**, 63–68 (2004).
- Kaneda, M. *et al. Nature* **429**, 900–903 (2004).
- www.ncbi.nlm.nih.gov/omim
- Oudejans, C. B. *et al. Mol. Hum. Reprod.* **10**, 589–598 (2004).

Structural basis for allostery in integrins and binding to fibrinogen-mimetic therapeutics

Tsan Xiao¹, Junichi Takagi^{1*}, Barry S. Collier², Jia-Huai Wang³ & Timothy A. Springer¹

¹The CBR Institute for Biomedical Research and Department of Pathology, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA

²Laboratory of Blood and Vascular Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

³Dana-Farber Cancer Institute, Departments of Pediatrics, Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 44 Binney Street, Boston, Massachusetts 02115, USA

* Present address: Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Integrins are important adhesion receptors in all Metazoa that transmit conformational change bidirectionally across the membrane. Integrin α and β subunits form a head and two long legs in the ectodomain and span the membrane. Here, we define with crystal structures the atomic basis for allosteric regulation of the conformation and affinity for ligand of the integrin ectodomain, and how fibrinogen-mimetic therapeutics bind to platelet integrin $\alpha_{IIb}\beta_3$. Allostery in the β_3 I domain alters three metal binding sites, associated loops and $\alpha 1$ - and $\alpha 7$ -helices. Piston-like displacement of the $\alpha 7$ -helix causes a 62° reorientation between the β_3 I and hybrid domains. Transmission through the rigidly connected plexin/semaphorin/integrin (PSI) domain in the upper β_3 leg causes a 70 Å separation between the knees of the α and β legs. Allostery in the head thus disrupts interaction between the legs in a previously described low-affinity bent integrin conformation, and leg extension positions the high-affinity head far above the cell surface.

Integrins are adhesion receptors that transmit signals bidirectionally across the plasma membrane^{1–4}. Rearrangements in integrin extracellular, transmembrane and cytoplasmic domains underlie diverse biological processes, including cell migration, morphogenesis, immune responses and vascular haemostasis. The platelet-specific integrin $\alpha_{IIb}\beta_3$ is important in both the arrest of bleeding at sites of vascular injury and pathological thrombosis leading to heart attacks and stroke. Loss of the vascular endothelium results in platelet deposition, and receptors for collagen, thrombin and other agonists then initiate platelet signalling, resulting in changes in the cytoplasmic domains of $\alpha_{IIb}\beta_3$ that are transmitted into conformational changes in the extracellular domains. This leads to high-affinity binding of fibrinogen and von Willebrand factor, resulting in crosslinking of platelets into aggregates by these multivalent ligands, and activation by $\alpha_{IIb}\beta_3$ of further intracellular signals. Mutations of either α_{IIb} or β_3 result in the bleeding disorder Glanzmann thrombasthenia and drugs that inhibit ligand binding to $\alpha_{IIb}\beta_3$ are effective in preventing and treating coronary artery thrombosis⁵.

Global structural rearrangements in integrin extracellular domains are demonstrated by electron microscopy and exposure of activation epitopes known as ligand-induced binding sites (LIBS)^{2,4}. Negative stain electron microscopy with image averaging of integrins has demonstrated three overall conformations of the extracellular domain^{3,6} (Fig. 1a–c). A low-affinity, bent conformation (Fig. 1a) matches an $\alpha_V\beta_3$ crystal structure^{7,8}. An extended form with a ‘closed’ headpiece conformation matching that in the crystal structure represents an intermediate affinity state (Fig. 1b). Ligand-binding induces a high-affinity, extended form with an ‘open’ headpiece, in which the angle between the β I and hybrid domains changes from acute to obtuse^{3,6} (Fig. 1c). This marked change in tertiary structure is supported by mutational studies^{3,6,9–11} and solution X-ray scattering¹². Ligand-mimetic compounds induce the extended, open headpiece conformation of integrins in solution and on the cell surface^{3,6,10–13}, and LIBS epitope exposure¹⁴. In contrast, when a ligand-mimetic is soaked into preformed

crystals containing the bent integrin conformation with the closed headpiece, binding induces only localized structural changes near the ligand binding site⁸.

In the low-affinity bent structure, the α and β subunit ectodomain carboxy termini⁷ and transmembrane domains are closely associated^{15,16}, and transmission of activation signals across the membrane involves separation between the α and β transmembrane and cytoplasmic domains^{16–18}. How allostery could be relayed between the integrin transmembrane domains, legs and ligand-binding head has been unclear. We have proposed that the conformation of the ligand-binding site atop the integrin β I domain could be transmitted to the outward swing of the β hybrid domain between the closed and open headpiece conformations (Fig. 1b, c) by a piston-like β I domain $\alpha 7$ -helix motion similar to that seen in integrin α I domains^{3,6}. However, in the absence of atomic views of the high-affinity integrin state, different opinions about its conformation have been put forward.

Here, atomic structures of $\alpha_{IIb}\beta_3$ fragments demonstrate the high-affinity, open conformation of the integrin headpiece, its binding to therapeutic antagonists, and the allosteric movements that link the ligand binding site of β I domains to $\alpha 7$ -helix displacement and outward swing of the hybrid domain. The β_3 hybrid and PSI domains act as a rigid lever that transmits and amplifies this motion, resulting in a 70 Å separation between the α and β legs at their knees that favours leg extension.

Overall structure of an open integrin headpiece

Two crystal forms each contain α_{IIb} residues 1–452 comprising the β -propeller domain and β_3 residues 1–440 including the β I, hybrid and PSI domains (Fig. 2a, b and Supplementary Table 1). Crystal form A contains one copy per asymmetric unit of the $\alpha_{IIb}\beta_3$ headpiece bound to 10E5 Fab¹⁹ (Fig. 2a), and ligand-mimetic antagonist or cacodylate ion bound to the β_3 I domain metal-ion-dependent adhesion site (MIDAS) (2.7–3.1 Å resolution). Crystal form B contains no Fab and three independent, cacodylate-bound $\alpha_{IIb}\beta_3$ molecules per asymmetric unit (2.9 Å resolution). The four

independent $\alpha_{IIB}\beta_3$ headpiece fragments from the two crystal forms adopt similar conformations, with only minor differences in the angle between the β I and hybrid domains (Fig. 2b). Five N-linked glycans are resolved, including one in a form A crystal lattice contact with seven carbohydrate residues (Fig. 2a). Residues 55–77 in our β_3 hybrid domain structures have a different sequence-structure register than previous β_3 structures (Supplementary Material).

Our crystal structures reveal an open, high-affinity conformation of the $\alpha_{IIB}\beta_3$ headpiece similar to that in electron microscopy averages of the $\alpha_V\beta_3$ ectodomain in which ligand binding induced the extended conformation with the open headpiece³ (Fig. 1c). Relative to the bent $\alpha_V\beta_3$ crystal structure with the closed, low-affinity conformation of the headpiece^{3,7}, the α 7-helix of the β_3 I domain moves downward, causing the hybrid and PSI domains to swing away from the α subunit by 62° (Fig. 2d). The angle between the β I and hybrid domains visualized here is in excellent agreement with that found by electron microscopy for liganded $\alpha_V\beta_3$ (ref. 3)

and the $\alpha_5\beta_1$ headpiece bound to FN3 module 10 of fibronectin⁶ (Fig. 2e). Ligand-mimetic antagonists are known to induce the conformation of $\alpha_{IIB}\beta_3$ with high affinity for fibrinogen¹⁴. Crystallization in the open conformation may also have been facilitated by truncation of the integrin tailpiece, which participates in extensive interfaces that stabilize the bent $\alpha_V\beta_3$ conformation³. The conformation characterized here termed ‘liganded-open’ was obtained by co-crystallization with ligand, enabling equilibration to the most favoured ligand-bound conformation before crystallization. By contrast, crystals of $\alpha_V\beta_3$ were previously formed in the bent, low-affinity conformation which we term ‘unliganded-closed’, and a ligand-mimetic antagonist was then soaked-in to obtain what we term the ‘liganded-closed’ conformation⁸. Crystal packing contacts are required for maintenance of the bent, liganded-closed conformation, because in solution addition of the same ligand to bent $\alpha_V\beta_3$ induces leg extension and conversion of the headpiece to the open conformation³.

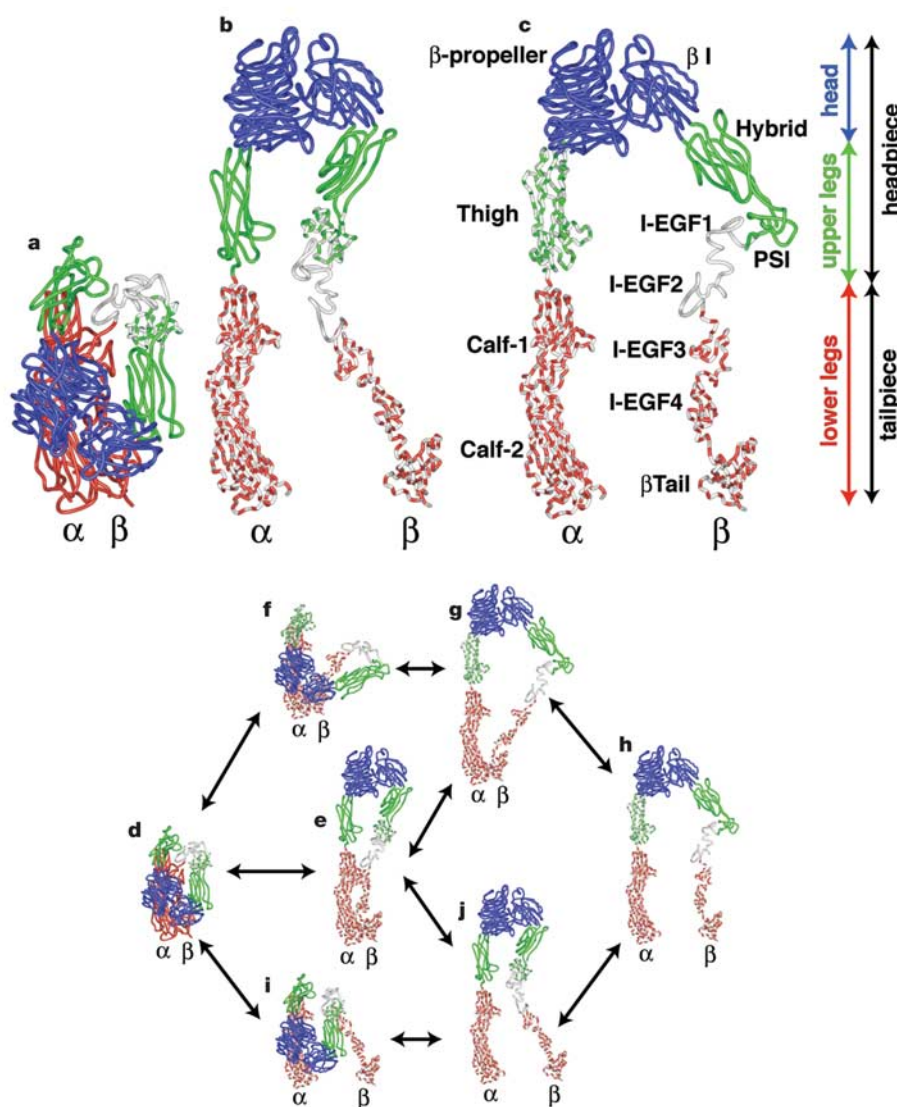


Figure 1 Quaternary rearrangements in the integrin ectodomain. **a–c**, Three conformational states visualized in electron microscopy^{3,6} and in crystal structures (here and in ref. 7). **d–j**, Proposed intermediates in equilibration between known conformational states. The upper pathways may be stimulated by ligand binding outside the cell, and the lower pathways by signals within the cell that separate the α and β subunit

transmembrane domains. Domains in **a–j** are shown in solid colour if known directly from crystal structures, dashed with grey if placed from crystal structures into electron microscopy image averages, and in solid grey for EGF-1 and EGF-2, which are modelled on EGF-3 and EGF-4.

Ligand binding to $\alpha_{IIb}\beta_3$

Small molecule antagonists bind to a small pocket atop the integrin head formed by loops from the α_{IIb} β -propeller and β_3 I domain (Figs 2a and 3a–d). The binding site for the macromolecular ligand fibrinogen is more extensive, as revealed by the positions of residues shown by mutation to affect binding (Fig. 3a). In macromolecular recognition, the β_3 specificity determining loop (SDL) and four α_{IIb} β -propeller loops that form a cap subdomain are particularly important, with 71% of fibrinogen-sensitive α_{IIb} mutations mapping to the latter²⁰. The α_{IIb} -specific 10E5 Fab binds solely to the α_{IIb} cap subdomain (Fig. 2a, Supplementary Material and Supplementary Fig. 2) and has no effect on its conformation (Fig. 2b). Binding of small molecules is not blocked by 10E5, as shown by their co-crystallization (Figs 2a and 3a–d), and thus blocks binding of fibrinogen to $\alpha_{IIb}\beta_3$ (ref. 19) solely by occluding the macromolecular recognition site on the α_{IIb} cap subdomain. By contrast, the β_3 specific-7E3 Fab, abciximab, blocks fibrinogen-

binding by binding to residues in the β_3 SDL^{5,21}.

The cap subdomain helps form the drug-binding pocket as well as the macromolecular recognition surface on α_{IIb} . It is comprised of four long insertions or loops in the β -propeller domain (Supplementary Fig. 1a). Insert 3 forms an α -helix in the α_{IIb} drug-binding pocket (Fig. 3a–d), and in integrin α subunits that contain I domains forms the ligand-binding α I domain⁴. Inserts 1 and 2 form a four-stranded anti-parallel β -sheet in the centre of the cap subdomain that is buttressed on one side by the α -helix and loop of insert 4, and on the other side by the β_3 SDL (Fig. 2a). Alternative splicing of cap insert 4 can regulate ligand-binding specificity^{22,23}.

The greatest structural differences between the α_{IIb} and α_V β -propellers are in cap inserts 1, 2 and 3 (Fig. 2d). The β_3 SDL closely associates with the cap subdomain, and structural variation between α_{IIb} and α_V appears to be responsible for the different β_3 SDL conformations in $\alpha_{IIb}\beta_3$ and $\alpha_V\beta_3$ (Figs 2d and 3d compared with 3f). By contrast, the remainder of the α_{IIb} and α_V β -propellers

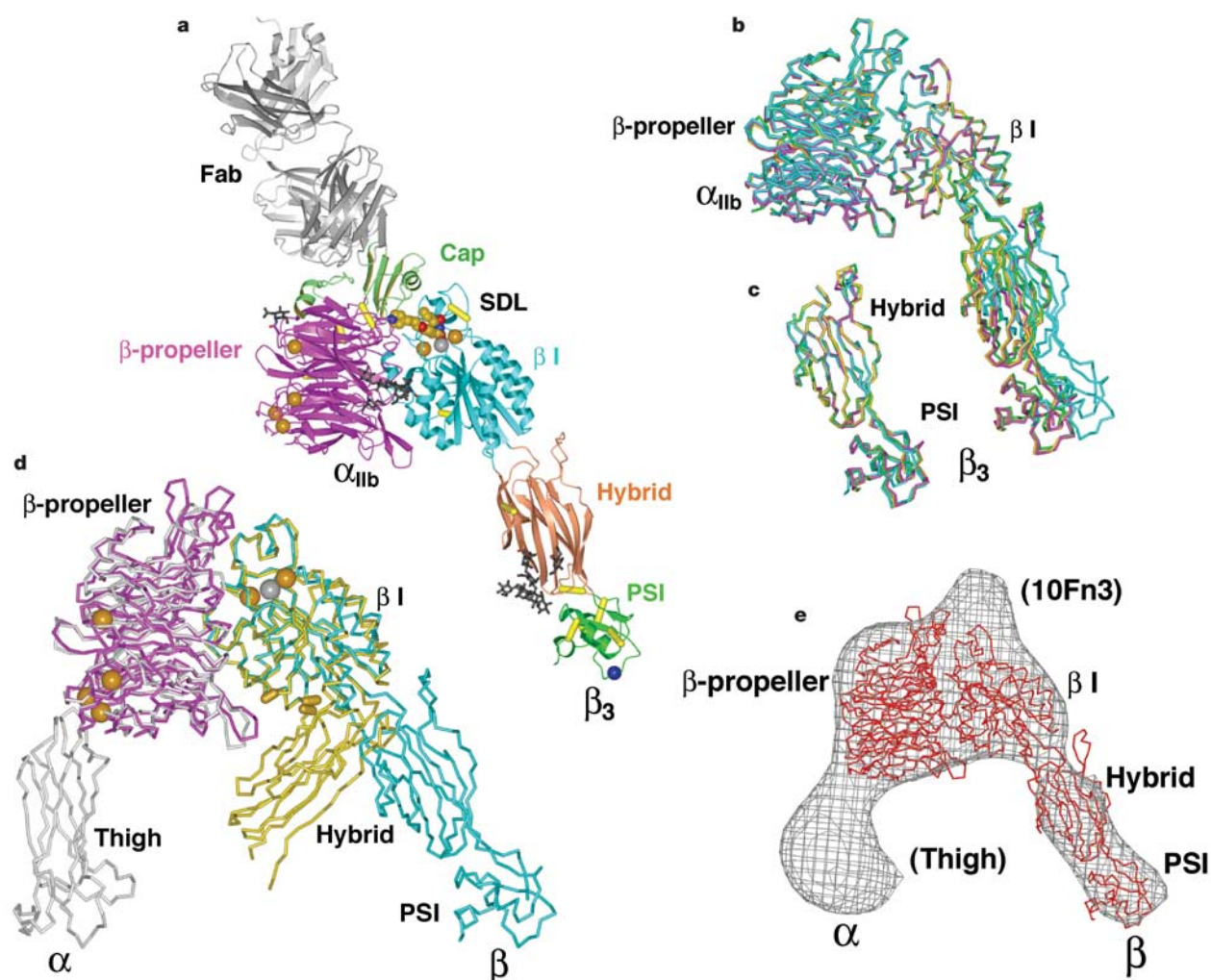


Figure 2 Structure of the $\alpha_{IIb}\beta_3$ headpiece. **a**, Ribbon diagram of the $\alpha_{IIb}\beta_3$:10E5 complex. Calcium and magnesium ions are shown as gold and silver spheres, respectively. Tirofiban is shown in cpk. The C α atom of HPA-1a alloantigenic determinant Leu 33 in the PSI domain is shown as a blue sphere. Glycan chains are displayed as black sticks. Integrin disulphides are shown as yellow C α –C α bonds. **b**, Superimposed on the basis of the β I domain are the three independent $\alpha_{IIb}\beta_3$ heterodimers in crystal form B (magenta, green and yellow), and one in form A (cyan). **c**, The hybrid and PSI domains of the four independent $\alpha_{IIb}\beta_3$ structures are superimposed. **d**, Liganded-open $\alpha_{IIb}\beta_3$ (crystal form A) and unliganded-closed $\alpha_V\beta_3$ headpieces⁷ are superimposed using the β I domain β -sheet. The α and β subunits are coloured magenta and cyan in $\alpha_{IIb}\beta_3$ and grey

and yellow in $\alpha_V\beta_3$. Calcium and magnesium ions in $\alpha_{IIb}\beta_3$ only are gold and silver spheres, respectively. Yellow cylinders in the β I/hybrid interface show positions of residues where introduction of N-glycosylation sites induces high affinity for ligand and LIBS epitope exposure^{9,42}. **e**, Superposition of an $\alpha_{IIb}\beta_3$ structure from crystal form B (red C α -trace) on the three-dimensional electron microscopy density (grey chickenwire) of the fibronectin-bound $\alpha_5\beta_1$ headpiece⁶. Domains are labelled and those only in the $\alpha_5\beta_1$ structure including the tenth FN3 domain of fibronectin are in parentheses. Figures in this paper utilize crystal form A unless stated otherwise and were prepared with programs Bobscript⁴⁶, Povray (The Povray Team, <http://www.povray.org>), Raster3D⁴⁷ and Ribbons⁴⁸.

are highly conserved structurally, with a root mean square deviation of 1.4 Å for 391 residues. The hub of the β -propeller that associates with the β I domain is especially conserved, enabling up to eleven diverse integrin α -subunits to bind to a single β -subunit²⁴. At their current resolutions of 2.7 Å and 3.1 Å, respectively, there is no significant difference between liganded-open $\alpha_{IIb}\beta_3$ and unliganded-closed $\alpha_v\beta_3$ in orientation between the β -propeller and β I domains (Fig. 2d).

All crystals reported here were formed in the presence of the physiological cations Ca^{2+} and Mg^{2+} . Three metal binding sites are present in the β_3 I domain in the drug-binding pocket⁸ (Figs 2a and 3a–f). In the middle site, the MIDAS, a Mg^{2+} ion coordinates a carboxyl group present in each of the three co-crystallized ligand-mimetic $\alpha_{IIb}\beta_3$ antagonists (Fig. 3a–d). The cacodylate ion that is bound to form A (Fig. 3e) and form B crystals binds in the same way as the carboxyl group of the ligand-mimetics, with one oxygen

coordinating the MIDAS Mg^{2+} and the other hydrogen-bonding to two backbone amides. The cacodylate may thus act as a pseudo-ligand, and stabilize the open β subunit I domain conformation similarly to pseudoligand lattice contacts that stabilize the open α subunit I domain conformation^{25–27}. Ca^{2+} was assigned at the two sites adjacent to the MIDAS on the basis of Yb^{3+} soaking data and coordination chemistry (see Methods). The site distal to the β -propeller is termed adjacent to MIDAS (ADMIDAS). The site near the β -propeller is termed ligand-induced metal binding site (LIMBS). The LIMBS is occupied in the open $\alpha_{IIb}\beta_3$ structure even when only a cacodylate pseudoligand is bound (Fig. 3e).

Two drugs used in prevention and treatment of coronary artery thrombosis, tirofiban and eptifibatide⁵, as well as Merck compound L-739758 (Fig. 3a–d), antagonize binding of fibrinogen to $\alpha_{IIb}\beta_3$. These compounds were developed as mimics of the Arg-Gly-Asp sequence that is found in a wide variety of integrin glycoprotein

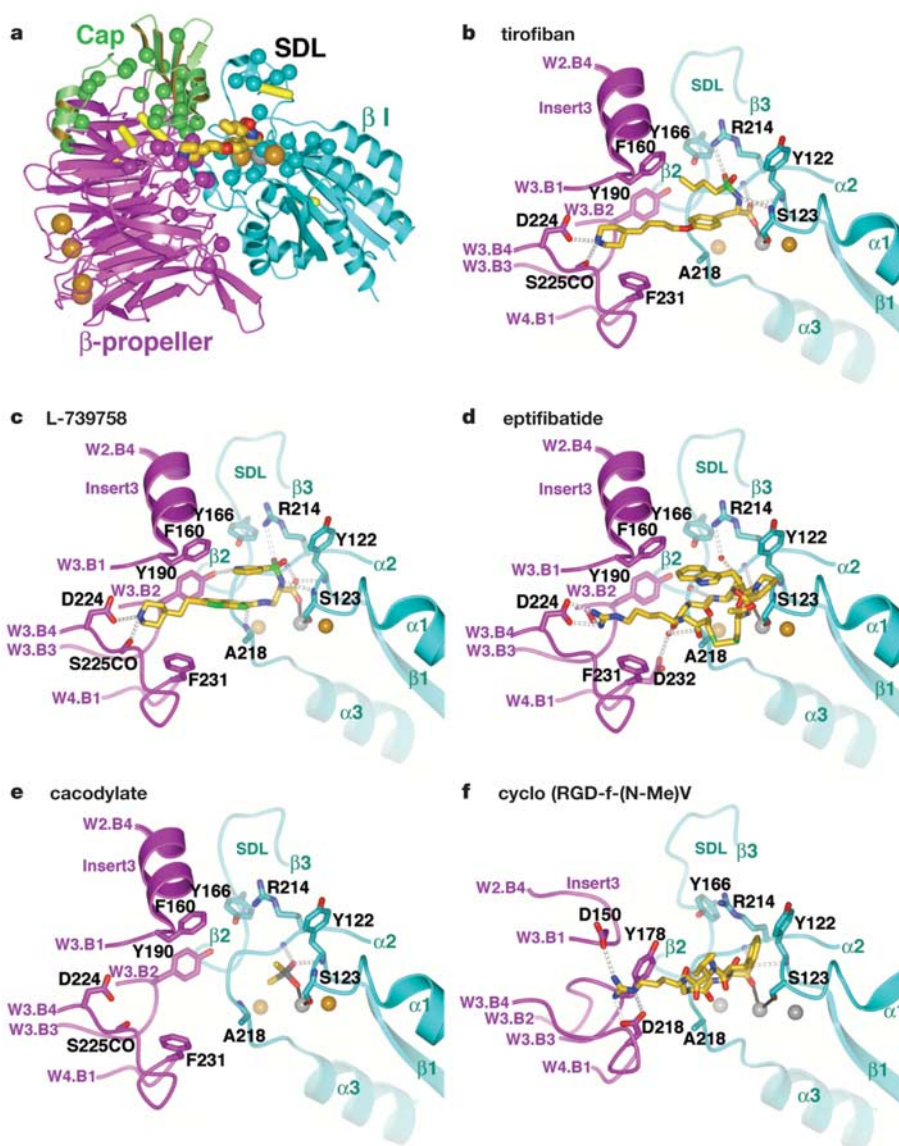


Figure 3 The binding sites for ligand-mimetic antagonists and fibrinogen at the α/β subunit interface. **a**, Mapping of fibrinogen binding sensitive mutations^{20,49,50} in $\alpha_{IIb}\beta_3$. Cp atoms of fibrinogen-binding sensitive residues are shown as spheres in the same colour as the domains in which they are present. The tirofiban-bound structure is shown. **b–f**, Binding of ligands or pseudoligands to $\alpha_{IIb}\beta_3$ (**b–e**) and binding of (**f**) cyclo Arg-Gly-Asp-D-Phe-N-methyl-Val (cyclo RGDfV) to $\alpha_v\beta_3$ (ref. 8). The orientation is

identical to that in **a**. The α and β subunits are shown in magenta and cyan, respectively. Small molecules are shown as ball-and-stick models with their carbon, nitrogen, oxygen, sulphur and arsenic atoms coloured yellow, blue, red, green and grey, respectively. Hydrogen bonds are shown as dotted lines. Ca^{2+} and Mg^{2+} ions are gold and silver spheres, respectively. The ligand and S123 coordinations to the MIDAS metal are shown as thin grey lines.

ligands, and the $\alpha_{IIb}\beta_3$ -binding sequence in fibrinogen, Lys-Gln-Ala-Gly-Asp-Val^{28,29,30}. In addition to $\alpha_{IIb}\beta_3$, the integrins $\alpha_V\beta_1$, $\alpha_V\beta_3$, $\alpha_V\beta_5$, $\alpha_V\beta_6$, $\alpha_V\beta_8$, $\alpha_5\beta_1$ and $\alpha_8\beta_1$ recognize the Arg-Gly-Asp motif. Therefore, a major focus of pharmaceutical development was selective inhibition of $\alpha_{IIb}\beta_3$, particularly in comparison to the closely related $\alpha_V\beta_3$ integrin. Our co-crystal structures (Fig. 3b–d), and comparison to an $\alpha_V\beta_3$ -selective compound soaked into $\alpha_V\beta_3$ crystals (Fig. 3f)⁸, reveal the basis for drug binding and selectivity (see Supplementary Material for details). Each drug has a basic moiety that mimics the arginine in Arg-Gly-Asp or the lysine in the fibrinogen sequence, and a carboxyl group that mimics the aspartic acid. Selectivity for $\alpha_{IIb}\beta_3$ and $\alpha_V\beta_3$ is conferred by a longer and shorter distance, respectively, between the basic and acidic moieties

of peptidomimetics (Fig. 3d, f), which can be adjusted by the constraints on the cyclic ring that bears the basic and aspartic acid residues, and the length of the basic residue side chain^{28,29}. Our structures show that this is because the basic ligand-mimetic side chain must reach further into the deeper β -propeller pocket of α_{IIb} to hydrogen-bond to α_{IIb} -Asp 224 (Fig. 3b–d), whereas in the $\alpha_V\beta$ -propeller the hydrogen-bonding residues Asp 150 and Asp 218 are nearer in its shallower pocket (Fig. 3f). Furthermore, Asp 218 in α_V is replaced by Phe 231 in α_{IIb} , favouring contacts with longer aliphatic moieties (Fig. 3b–d, f).

A large number of snake venom disintegrins with Arg-Gly-Asp sequences have evolved to inhibit haemostasis but are not $\alpha_{IIb}\beta_3$ -selective. An unusual, $\alpha_{IIb}\beta_3$ -selective disintegrin with a Lys-Gly-

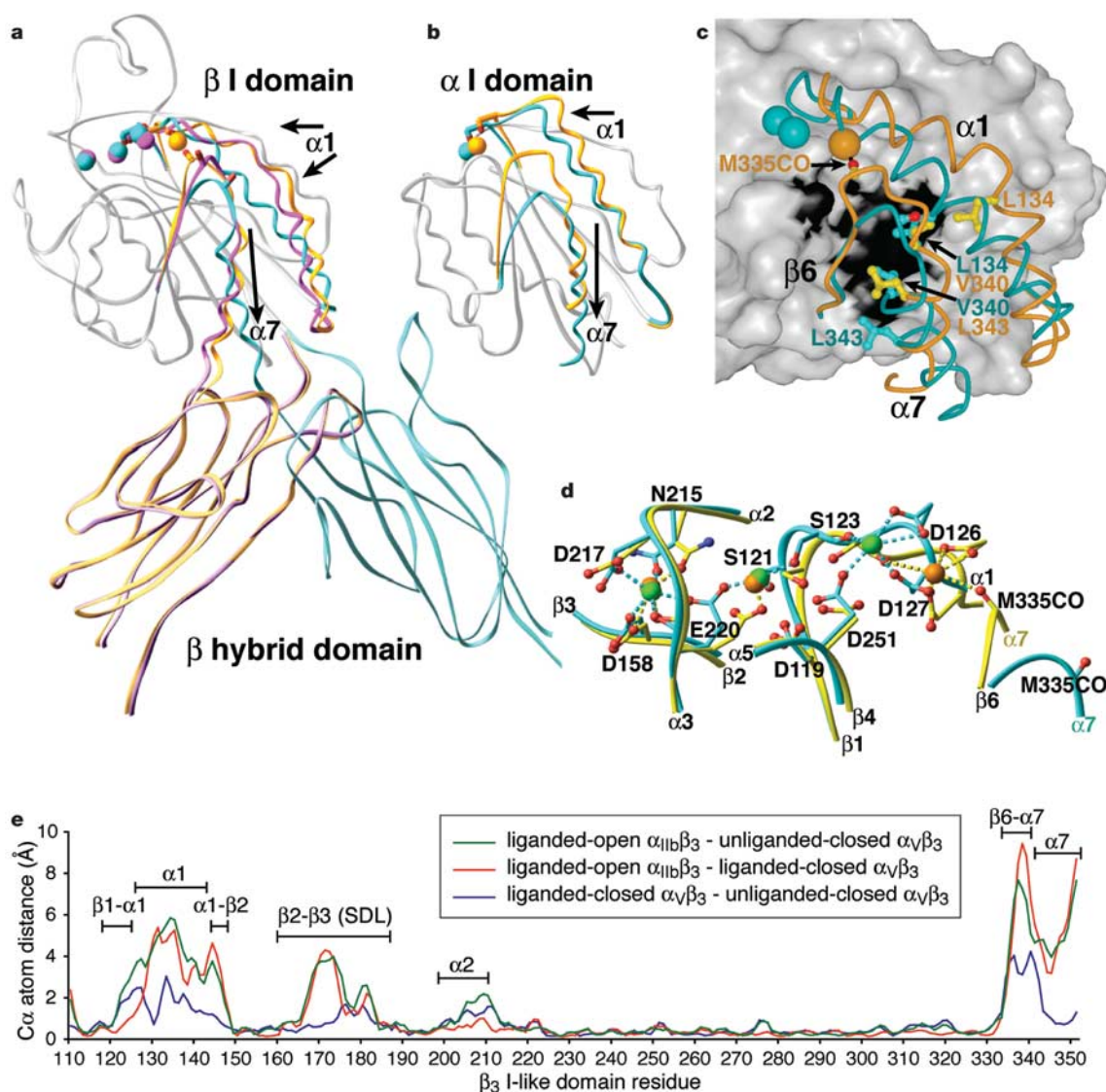


Figure 4 Allostery in the β I domain and comparison with α I domain. **a**, Overview of motions in the β_3 I and hybrid domains. Non-moving parts of the backbone are shown as a grey worm. Moving segments shown as C α -traces are from unliganded-closed $\alpha_V\beta_3$ (gold), liganded-closed $\alpha_V\beta_3$ (magenta) and liganded-open $\alpha_{IIb}\beta_3$ (cyan). The direction of movement is shown with arrows. **b**, Comparison with α I domains. The moving segments of unliganded-closed (gold) and pseudoliganded-open (cyan) α_M I domains^{25,26} and their MIDAS metal ions are shown as in **a** and in the same orientation. **c**, Hydrophobic ratchet pockets underlying the $\beta 6$ - $\alpha 7$ loop and $\alpha 1$ -helix. The unliganded-closed (orange) and liganded open (cyan) $\beta 1$ - $\alpha 1$ loop, $\alpha 1$ - $\beta 2$ loop and $\beta 6$ - $\alpha 7$ loop and $\alpha 7$ -helix are shown as worm traces with key side chains, the Met 335 carbonyl and metal ions in

the same colour. The rest of the domain is shown as a grey surface, except for hydrophobic pocket residues Tyr 116, Val 247, Thr 249, Ile 307, Ala 309 and Thr 311, which are shown as a black surface. **d**, β_3 I-like domain metal coordination sites in liganded-open $\alpha_{IIb}\beta_3$ (cyan) and unliganded-closed $\alpha_V\beta_3$ (yellow). LIMBS, MIDAS and ADMIDAS positions are shown left to right in similar orientation as in **a**. The LIMBS and MIDAS were not occupied in the unliganded-closed structure⁸; for reference, metal ions at these sites are shown from the liganded-closed structure⁸. In **a**–**d**, metal ions are shown as spheres in the same or a similar (**d**) colour as their associated backbone. **e**, Distances between C α atoms in the three superimposed β I domains, smoothed by averaging at each residue over a 3-residue window.

Asp-Trp sequence led to the development of eptifibatide³¹. The lysine with its longer aliphatic side chain than arginine confers selectivity, whereas the tryptophan confers high affinity. In an independent drug development programme, exosite substituents such as pyridyl sulphonamide were found that substantially increase affinity³⁰ (Fig. 3c) and that occupy the same position in the drug-binding pocket as the tryptophan residue (Fig. 3d).

Comparison to the cacodylate-bound structure (Fig. 3e) demonstrates that the drug-binding pocket in $\alpha_{\text{IIB}}\beta_3$ is rigid, with the contacting residues adopting the same conformation with or without the drugs. This helps account for high affinity, despite burial of a solvent-accessible surface area of only 300–390 Å² on the three drugs.

An open, high-affinity conformation for integrin β_3 I domains

Comparisons among unliganded-closed $\alpha_{\text{V}}\beta_3$, liganded-closed $\alpha_{\text{V}}\beta_3$ and liganded-open $\alpha_{\text{IIB}}\beta_3$ reveal the atomic basis for communication of allostery between the ligand binding site in the β_3 I domain and other integrin domains (Figs 2d and 4a, c, d, e; Supplementary Movie 1). In liganded-open $\alpha_{\text{IIB}}\beta_3$ compared with unliganded-closed $\alpha_{\text{V}}\beta_3$, concerted movements in the β_3 I domain occur in the ligand-binding β_1 – α_1 loop, the α_1 -helix, the α_1 – β_2 loop, the α_2 -helix, the β_6 – α_7 loop and the α_7 -helix. Movements in the ligand- and metal-binding β_1 – α_1 loop are similar in both liganded structures, but greater in magnitude for the liganded-open structure than the liganded-closed structure (2.9 and 2.0 Å per residue for residues 121–127, respectively) (Fig. 4a, e). A single turn

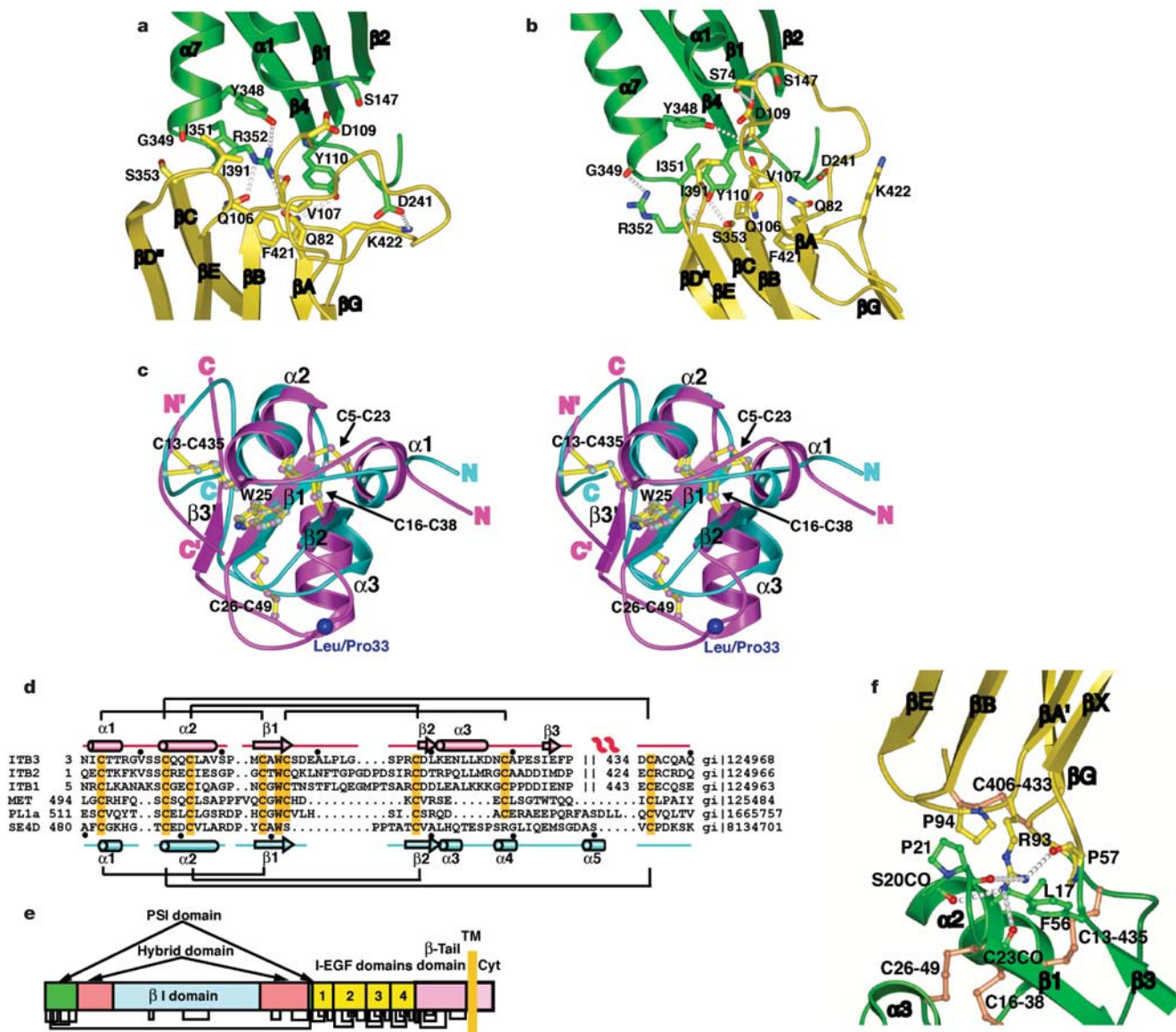


Figure 5 The hybrid and PSI domains and their interfaces. **a**, **b**, The β_3 I/hybrid domain interface in the unliganded-closed structure (**a**) and liganded-open structure (**b**). Ribbon backbone and side-chain carbon atoms are shown in green (β_3 I) and yellow (hybrid) with the β_3 I domain β -sheet in the same orientation. The α_7 -helical ribbon is shown up to the same residue (350) in both structures to aid comparison of α_7 -helix position. **c**, Stereo view of the superposition of the PSI domains from β_3 (magenta) and semaphorin4D (cyan). The disulphide bridges and the conserved tryptophan are shown as ball-and-stick models with their bonds coloured yellow and atoms the same colour as the backbone. The Leu/Pro33 alloantigen site is represented with a large blue C α sphere. The amino and

carboxyl termini of each domain are indicated. The N' and C' refer to termini for residues 434–440 that constitute part of the PSI and EGF-1 domains. **d**, Sequence alignment of PSI domains from integrins, semaphorin4D (SE4D), a plexin and c-met. Disulphide connections are shown above (β_3) and below (semaphorin4D) the respective sequences. The conserved cysteines and tryptophans are highlighted in orange. The secondary structures are shown for β_3 (top) and sema4D (bottom). **e**, Domain organization of integrin β subunits, showing multiple domain insertions. The revised disulphide bond pattern is shown below. **f**, The interface between the hybrid (yellow) and PSI (green) domains. Disulphide bonds are shown in orange.

of 3_{10} -helix in this loop bearing the key MIDAS and ADMIDAS residue Ser 123 moves *en bloc* in the liganded-open structure, but the lesser movement in the liganded-closed structure disrupts this 3_{10} -helix.

Extensive movements occur in the $\beta 6$ - $\alpha 7$ loop and the $\alpha 7$ -helix between the liganded-open and unliganded-closed structures. The $\beta 6$ - $\alpha 7$ loop moves downward, and the $\alpha 7$ -helix moves downward and pivots laterally, resulting in an average displacement for residues 333–352 of 5.3 Å (Fig. 4a, c, e). Between the unliganded-closed and liganded-closed structures there is a movement of lesser magnitude in the $\beta 6$ - $\alpha 7$ loop; however, the movements differ in direction (Fig. 4a, e). Movement in this region of the liganded-closed structure is therefore a consequence of strain rather than movement along the pathway towards the open conformation.

In the liganded-open compared with the unliganded-closed structure, the $\alpha 1$ -helix moves³² upward to accommodate movements near its beginning at the MIDAS and ADMIDAS, and near its end, that allow hybrid domain swing-out (Fig. 4a). Furthermore, a bend between the 3_{10} -helix in the $\beta 1$ - $\alpha 1$ loop and the $\alpha 1$ -helix is straightened, and the $\alpha 1$ -helix is lengthened by five residues, as it pivots laterally to fill-in room vacated by the $\beta 6$ - $\alpha 7$ loop (Fig. 4a, c). The position of the $\alpha 1$ -helix in the liganded-closed structure does not resemble that in the unliganded-closed or the liganded-open structures (Fig. 4a, e), suggesting that its position is off the pathway towards the open conformation, and results from strain imposed by binding to ligand when the interface with the hybrid domain remains in the closed conformation.

Changes in metal ion coordination are closely related to loop movements in the β I domain, and are key for stabilizing its alternative conformational states (Fig. 4d). Coordination of the Met 335 backbone carbonyl in the $\beta 6$ - $\alpha 7$ loop to the ADMIDAS Ca^{2+} ion in the unliganded-closed conformation (Fig. 4d)⁷ is broken in liganded-closed and Mn^{2+} -bound-closed $\alpha_V\beta_3$ ⁸, and in liganded-open $\alpha_{IIb}\beta_3$ (Fig. 4d). Mn^{2+} activates integrins by competing with Ca^{2+} at the ADMIDAS; the much lower propensity of Mn^{2+} than Ca^{2+} for carbonyl coordination enables downward displacement of the $\beta 6$ - $\alpha 7$ loop and activation in Mn^{2+} (ref. 33). Breakage of the Met 335 coordination and the movement of the $\beta 1$ - $\alpha 1$ loop with its coordinating residues enable the large movement in position of the ADMIDAS metal between the unliganded-closed and liganded-open conformations (Fig. 4d). Additionally, small shifts in the position of the MIDAS and LIMBS metals relate to marked movements and changes in metal ion coordination of residues Asp 251, Glu 220, Asn 215 and Asp 217 (Fig. 4d). The metal ions seem to occupy similar but less strained orientations in the liganded-open compared with the liganded-closed structures, with the caveat that the resolutions are 2.7 and 3.3 Å, respectively. For example, the Asp 217 side chain coordinates with the LIMBS metal ion in the liganded-open structure, instead of orienting away from it in the liganded-closed structure^{8,33} (Fig. 4d). Its position in the liganded-open structure is more consistent with the function of Asp 217 in $\beta 7$ as a LIMBS residue that stabilizes the high-affinity state³³.

Rearrangements in the β I domain are clearly structurally homologous to those in α I domains (Fig. 4a, b). Movements of similar directionality occur in the MIDAS metal ion, $\beta 1$ - $\alpha 1$ loop, $\alpha 1$ -helix, $\beta 6$ - $\alpha 7$ loop and $\alpha 7$ -helix⁴. Hydrophobic ratchet residues that are located one turn of 3_{10} -helix apart, stabilize alternative $\beta 6$ - $\alpha 7$ loop positions in both α I and β I domains. In contrast to the one-turn displacement in the intermediate²⁷ and two-turn displacement in the open α I domain conformations^{25,26} (Fig. 4b), a one-turn displacement occurs in the open β I domain conformation (Fig. 4a, c). Val 340, located in an upper hydrophobic pocket in the closed conformation, moves to a lower hydrophobic pocket in the open conformation, displacing Leu 343 (Fig. 4c). Movement of the $\alpha 1$ -helix plays a similar part in activation of α I and β I domains by accompanying the movement of metal-coordinating

residues in the $\beta 1$ - $\alpha 1$ loop; however, the greater magnitude of $\alpha 1$ -helix movement in the β I domain is accompanied by a unique ratchet-like movement: residue Leu 134 in the $\alpha 1$ -helix moves laterally to occupy the upper hydrophobic pocket vacated by Val 340 (Fig. 4b).

Hybrid domain swing-out

The piston-like displacement of the β I domain $\alpha 7$ -helix, which connects to the hybrid domain β C strand, results in complete remodelling of the interface between these domains (Figs 4a and 5a, b; Supplementary Movie 1). The interface in the closed conformation covers 1,350 Å² and includes at its centre hydrogen-bonding residues Tyr 110, Tyr 348 and Arg 352⁷. Upon opening not only does the $\alpha 7$ -helix move downward, but the last two residues unwrap from the helix, including Arg 352, which reorients to the exterior of the interface. The more extended conformation at the junction between $\alpha 7$ and β C, and the straightening of the angle between $\alpha 7$ and β C, augment the effect of $\alpha 7$ displacement on hybrid domain swing-out. The more extended, end-to-end orientation between the β I and hybrid domains results in a smaller interface of about 800 Å² (Figs 4a and 5a, b). Near its centre the interface contains Tyr 110 and Tyr 348, which adopt new hydrogen-bonding partners (Fig. 5b). Reorganization of hydrogen-bonded interfaces is the general mechanism for allosteric transitions³⁴. The smaller size of the open β I/hybrid interface allows some flexibility. Relative to the closed conformation, the hybrid domain swings out 69° in crystal form A, and 57°, 59° and 61° in the three molecules in crystal form B (Fig. 2b).

Structure of an integrin PSI domain

The structure of a PSI domain in an integrin (Fig. 2a and Supplementary Fig. 3b) and comparison to that in semaphorin 4D³⁵ demonstrate the predicted homology of these domains in plexins, semaphorins and integrins³⁶, and an unexpected insertion. The buried $\beta 1$ strand bearing the invariant tryptophan, and the disulphide-bonded $\alpha 1$ - and $\alpha 2$ -helices are well conserved, but other regions differ significantly, including the longer $\alpha 3$ -helix (Fig. 5c, d). All three shared disulphide bonds superimpose well, and a fourth shared with integrins, plexins and the growth factor receptor MET, but not semaphorins, is also revealed (Fig. 5c, d). Chemical assignment of disulphides in the integrin β_3 PSI domain was difficult because cysteines are so closely spaced in sequence³⁷; the structural data reassigns all of these disulphides, including the long-range disulphide, which was previously identified as Cys 5 to Cys 435, and is now shown to link Cys 13 to Cys 435 (Fig. 5d, e and Supplementary Fig. 3b). The long-range disulphide superimposes well with the highly conserved intradomain disulphide that links the second cysteine to the last cysteine in the PSI domain of plexins and semaphorins (Fig. 5c), and this second cysteine aligns perfectly in all PSI domains, maintaining its relationship to the third cysteine in the conserved $\alpha 2$ -helix (Fig. 5d). These findings suggest that β_3 Cys 435 is the eighth cysteine of the PSI domain, and an integral part of the PSI domain fold. Therefore, there seems to be a nested domain insertion in the β_3 structure: the β I domain is inserted in the hybrid domain, which is in turn inserted in the PSI domain (Fig. 5e).

Superposition of four independent molecules shows that the 860 Å² interface between the hybrid and PSI domain is rigid (Fig. 2c). Arg 93 of the hybrid domain, which is invariant in vertebrate integrins, inserts its side chain into the centre of the interface, making hydrogen bonds to one hybrid domain residue and three different PSI domain residues (Fig. 5f). Rigidity of the interface is further supported by the nearby Cys 13-Cys 435, Cys 16-Cys 38 and Cys 26-Cys 49 disulphides in the PSI domain, and the Cys 406-Cys 433 disulphide in the hybrid domain (Fig. 5f). The connection to integrin EGF-like (I-EGF) domain 1 may also be rigid, because a portion of it containing residues 437–440, including

Cys 437 that is predicted to be disulphide-linked within I-EGF1³⁸, is present in the structure and stabilized by main-chain-side-chain and side-chain-side-chain hydrogen bonds with the PSI domain.

The human platelet alloantigen HPA-1 or Pl^A system, of great importance for neonatal alloimmune thrombocytopenia and post-transfusion purpura, corresponds to a leucine/proline polymorphism at PSI residue 33 (ref. 39). Our structure shows that the Leu 33 side chain is well exposed to solvent (Fig. 2a) on a loop of the PSI domain that is flexible and particularly long in integrins (Fig. 5c, d). Polymorphic substitution of the distally located residue Arg 93 at the hybrid/PSI interface (Fig. 5f) disrupts the HPA-1a epitope⁴⁰, demonstrating the importance of the interface for structural integrity of the PSI domain. The I-EGF1 domain also participates in the HPA-1 epitope⁴¹, further emphasizing the tight linkage between the hybrid, PSI and I-EGF1 domains.

Allosteric mechanism for integrin activation

The structural rearrangements demonstrated here between the closed and open conformations seem to be general for all integrins. The effects of mutations designed to induce hybrid domain swing-out^{9,42} (Fig. 2d), allosteric activating or inhibitory mAb that bind to the hybrid domain^{10,13} (Supplementary Fig. 4), disulphide cross-links in the $\beta 6$ - $\alpha 7$ loop⁴³ and shortening of the $\alpha 7$ -helix in the β I domain⁴⁴, all support the conclusion that the closed and open conformations of the integrin headpiece have low and high affinity for ligand, respectively. Furthermore, these experiments have been conducted on β_1 , β_2 , β_3 and β_7 integrins, demonstrating the generality of our findings. Moreover, the structural rearrangements shown here are consistent with exposure of LIBS and activation epitopes, including on the PSI domain in models of the complete integrin ectodomain (Supplementary Materials).

Our study reveals the β I/hybrid domain interface as the epicentre for quaternary structural rearrangements in integrins. A movement of about 10 Å occurs at the junction between the β I domain $\alpha 7$ -helix and the hybrid domain β C-strand that results in a 62° pivot at the β I/hybrid domain interface. The rigidly connected hybrid and PSI domains act as a mechanical lever in the upper β leg that amplifies and transmits β I domain allostery to the knee between the upper and lower β legs, resulting in a 70 Å displacement at the PSI domain. Because the PSI domain is near the β knee between I-EGF1 and I-EGF2, and in the closed conformation of the headpiece the β knee is near the α subunit knee or genu, the two legs must separate by about 70 Å at their knees (Fig. 1b, c). Swing-out of the upper β leg could readily occur if it were preceded by extension at the α and β knees (Fig. 1d to 1e to 1g). Electron microscopy studies show that below the PSI domain, the β leg is flexible in the extended conformation, whereas when the α leg extends, it adopts a single favoured orientation³. Interestingly, swing-out of the upper β leg might also occur in the bent conformation (Fig. 1d to 1f), because if the β subunit upper and lower legs moved as a rigid body, there would be no clash with the α subunit, facilitating conformational equilibration, and despite the 70 Å separation at the knees, the C termini of the α and β subunit ectodomains would only move apart by 15 Å. Thus, adjustments in the flexible β subunit lower leg domains could allow the α and β subunit transmembrane domains to remain associated during upper leg swing-out in the bent conformation (Fig. 1d to 1f to 1g). Transmembrane domain separation could thus occur as a later event in the process of integrin activation by ligand from outside the cell (Fig. 1g to 1h), whereas transmembrane domain separation is a key early step in activation by signals from within the cell^{3,16} (Fig. 1d to 1e to 1j to 1h, or Fig. 1d to 1i to 1j to 1h). In the bent $\alpha_v\beta_3$ integrin conformation, there are large, hydrophilic interfaces of 2,000 Å² each between the headpiece and the legs, and between the α and β subunit legs³. The latter interface would be disrupted by upper β leg swing-out, thereby weakening that between the headpiece and the legs, and facilitating adoption of the extended integrin conformation with the open

headpiece observed for liganded integrins by electron microscopy^{3,6} (Fig. 1c and 1h).

Our crystal structures, together with previous structural studies on integrins^{3,6-8,12,15,38}, now provide a clear picture of the conformational rearrangements in the integrin head that regulate affinity for ligand, and how conformational signals are transmitted to the leg domains. Further structures are needed to define I-EGF domains 1 and 2 in both the bent and extended conformations, the apparently unique conformation of the α subunit genu in the extended conformation³, and how allostery is communicated between the β I and α I domains⁴. □

Methods

Protein expression, purification and crystallization

For details and references, see Supplementary Materials. Briefly, the soluble $\alpha_{IIb}\beta_3$ headpiece encompassing residues 1–621 of α_{IIb} and residues 1–472 of β_3 was expressed in CHO Lec 3.2.8.1 cells with an ACID-BASE coiled-coil clasp at the C termini, as described for soluble $\alpha_5\beta_1$ ⁴⁵, except that a hexahistidine tag was fused to the C terminus of β_3 . The expressed protein was purified by ammonium sulphate precipitation, Ni-NTA affinity chromatography and size exclusion chromatography (Superdex 200 HR), concentrated to 1 mg ml⁻¹, and treated with chymotrypsin at 25 °C for 16 h. The unclashed (coiled-coil and His₆ tag removed) $\alpha_{IIb}\beta_3$ protein was collected in the flow-through of a second Ni-NTA chromatography step. The purified $\alpha_{IIb}\beta_3$ was mixed with the 10E5 Fab (1:1.1 molar ratio) and the complex was purified by Superdex 200 chromatography. The complex was subjected to digestion with carboxypeptidase A and B (Calbiochem) (1:100 weight ratio) in the presence of 1 mM ZnCl₂ at 25 °C for 16 h. A stable protease resistant core of $\alpha_{IIb}\beta_3$ was obtained and further purified by a final Superdex 200 chromatography step and stored in TBS plus calcium and magnesium, and used to obtain crystal form A. To obtain crystal form B, $\alpha_{IIb}\beta_3$ fragment purified through the second Ni-NTA chromatography step was mixed with an excess of the purified fibrinogen γ chain C-terminal domain fragment (see Supplementary Information) in the presence of 1 mM MnCl₂ and subjected to carboxypeptidase A and B treatment. This resulted in the same pattern of $\alpha_{IIb}\beta_3$ digestion as for the 10E5 Fab complex sample; however, little of the fibrinogen domain co-purified with the $\alpha_{IIb}\beta_3$ headpiece upon Superdex 200 chromatography, probably owing to hydrolysis of the $\alpha_{IIb}\beta_3$ -binding C-terminal residues of the fibrinogen γ chain by carboxypeptidase. This material was used to obtain crystal form B, which contains no fibrinogen fragment.

The Topaz crystallizer (Fluidigm) was used to identify initial crystallization conditions by free interfacial diffusion and the lead conditions were optimized with hanging-drop vapour diffusion. The final optimized well solution for form A crystals of the 10E5 Fab: $\alpha_{IIb}\beta_3$ complex is 11% PEG 3350, 0.7 M magnesium acetate and 0.1 M sodium cacodylate, pH 6.5, and for crystal form B is 10% PEG 8000, 0.4 M magnesium acetate and 0.1 M sodium cacodylate, pH 7.0. Acetate and 4 °C temperature were absolutely required for crystallization. To obtain co-crystals with the drugs, protein sample was mixed with each drug at 1:3 to 1:5 molar ratios before setting up the hanging drops. The optimized crystallization conditions were 10–12% PEG 3350, 0.7 M magnesium acetate and 0.1 M imidazole (pH 6.5) in place of cacodylate.

Structure determination

Diffraction data were collected at the 19-ID station of the Advanced Photon Source (APS) at the Argonne National Laboratory and the A-1 station of the Cornell High Energy Synchrotron Source (CHESS). The structure of crystal form A was determined by molecular replacement using search models of the β I domain and the β I domain plus the β -propeller from $\alpha_5\beta_1$ (PDB ID 1L5G), and a murine Fab 36–71 (PDB ID 6FAB). Electron density maps calculated using phases from the search models clearly showed the presence of the hybrid domain, plus difference densities in the CDR loops of the Fab. The hybrid domain was placed in density and rebuilt. Excellent densities for landmark residues Phe 56, Pro 57, Pro 68 and Leu 69 in the first β -strand of the hybrid domain and multiple nearby disulphide bonds in the hybrid and PSI domains (Fig. 5f and Supplementary Fig. 3a) necessitated a change in the sequence-to-structure register of this β -strand. The structure of the PSI domain was built on the basis of the electron density maps computed with refined models and guided by the secondary structure arrangements in the sema4D crystal structure³⁵. Strong electron density for the conserved disulphides, the single tryptophan and the secondary structures allowed unambiguous tracing of the whole domain. Continuous electron density extends beyond residue Cys 435 and allowed the building of residues Ala 436 to Gln 440 of I-EGF1, although alternative conformations of these residues cannot be ruled out. The structure of crystal form B was solved by molecular replacement using the $\alpha_{IIb}\beta_3$ structure in crystal form A as a search model. The structures of the drug bound $\alpha_{IIb}\beta_3$ were solved by molecular replacement using the native structure as a search model. Electron density for the bound drug molecules was readily discernible in the maps calculated with the molecular replacement solution.

Received 25 June; accepted 26 August 2004; doi:10.1038/nature02976.

Published online 19 September 2004; corrected 6 October 2004 (see full text online).

1. Hughes, P. E. & Pfaff, M. Integrin affinity modulation. *Trends Cell Biol.* **8**, 359–364 (1998).
2. Takagi, J. & Springer, T. A. Integrin activation and structural rearrangement. *Immunol. Rev.* **186**, 141–163 (2002).
3. Takagi, J., Petre, B. M., Walz, T. & Springer, T. A. Global conformational rearrangements in integrin extracellular domains in outside-in and inside-out signaling. *Cell* **110**, 599–611 (2002).

4. Springer, T. A. & Wang, J.-h. in *Cell Surface Receptors* (ed. Garcia, K. C.) (Elsevier, San Diego, 2004).
5. Collier, B. S. Platelet GPIIb/IIIa antagonists: the first anti-integrin receptor therapeutics. *J. Clin. Invest.* **99**, 1467–1471 (1997).
6. Takagi, J., Strokovich, K., Springer, T. A. & Walz, T. Structure of integrin $\alpha 5\beta 1$ in complex with fibronectin. *EMBO J.* **22**, 4607–4615 (2003).
7. Xiong, J.-P. *et al.* Crystal structure of the extracellular segment of integrin $\alpha V\beta 3$. *Science* **294**, 339–345 (2001).
8. Xiong, J. P. *et al.* Crystal structure of the extracellular segment of integrin $\alpha V\beta 3$ in complex with an Arg-Gly-Asp ligand. *Science* **296**, 151–155 (2002).
9. Luo, B.-H., Springer, T. A. & Takagi, J. Stabilizing the open conformation of the integrin headpiece with a glycan wedge increases affinity for ligand. *Proc. Natl Acad. Sci. USA* **100**, 2403–2408 (2003).
10. Luo, B.-H., Strokovich, K., Walz, T., Springer, T. A. & Takagi, J. Allosteric $\beta 1$ integrin antibodies that stabilize the low affinity state by preventing the swing-out of the hybrid domain. *J. Biol. Chem.* **279**, 27466–27471 (2004).
11. Luo, B.-H., Springer, T. A. & Takagi, J. High affinity ligand binding by integrins does not involve head separation. *J. Biol. Chem.* **278**, 17185–17189 (2003).
12. Mould, A. P. *et al.* Structure of an integrin-ligand complex deduced from solution X-ray scattering and site-directed mutagenesis. *J. Biol. Chem.* **278**, 39993–39999 (2003).
13. Mould, A. P. *et al.* Conformational changes in the integrin βA domain provide a mechanism for signal transduction via hybrid domain movement. *J. Biol. Chem.* **278**, 17028–17035 (2003).
14. Du, X. *et al.* Ligands “activate” integrin $\alpha IIb\beta 3$ (platelet GPIIb-IIIa). *Cell* **65**, 409–416 (1991).
15. Adair, B. D. & Yeager, M. Three-dimensional model of the human platelet integrin $\alpha IIb\beta 3$ based on electron cryomicroscopy and X-ray crystallography. *Proc. Natl Acad. Sci. USA* **99**, 14059–14064 (2002).
16. Luo, B.-H., Springer, T. A. & Takagi, J. A specific interface between integrin transmembrane helices and affinity for ligand. *PLoS Biol.* **2**, 776–786 (2004).
17. Vinogradova, O. *et al.* A structural mechanism of integrin $\alpha IIb\beta 3$ “inside-out” activation as regulated by its cytoplasmic face. *Cell* **110**, 587–597 (2002).
18. Kim, M., Carman, C. V. & Springer, T. A. Bidirectional transmembrane signaling by cytoplasmic domain separation in integrins. *Science* **301**, 1720–1725 (2003).
19. Collier, B. S., Peerschke, E. I., Scudder, L. E. & Sullivan, C. A. A murine monoclonal antibody that completely blocks the binding of fibrinogen to platelets produces a thrombasthenic-like state in normal platelets and binds to glycoproteins IIb and/or IIIa. *J. Clin. Invest.* **72**, 325–338 (1983).
20. Kamata, T., Tieu, K. K., Springer, T. A. & Takada, Y. Amino acid residues in the αIIb subunit that are critical for ligand binding to integrin $\alpha IIb\beta 3$ are clustered in the β -propeller model. *J. Biol. Chem.* **276**, 44275–44283 (2001).
21. Artoni, A. *et al.* The specificity determining loop and α helix 1 on human integrin $\beta 3$ determine the binding of murine monoclonal antibody 7E3 to $\alpha IIb\beta 3$: implications for the mechanism of integrin activation. *Proc. Natl Acad. Sci. USA* (in the press) (2004).
22. Zavortink, M., Bunch, T. A. & Brower, D. L. Functional properties of alternatively spliced forms of the *Drosophila* PS2 integrin α subunit. *Cell Adhes. Commun.* **1**, 251–264 (1993).
23. von der Mark, H. *et al.* Alternative splice variants of $\alpha 7\beta 1$ integrin selectivity recognize different laminin isoforms. *J. Biol. Chem.* **277**, 6012–6016 (2002).
24. Springer, T. A. Predicted and experimental structures of integrins and β -propellers. *Curr. Opin. Struct. Biol.* **12**, 802–813 (2002).
25. Lee, J.-O., Rieu, P., Arnaout, M. A. & Liddington, R. Crystal structure of the A domain from the α subunit of integrin CR3 (CD11b/CD18). *Cell* **80**, 631–638 (1995).
26. Lee, J.-O., Bankston, L. A., Arnaout, M. A. & Liddington, R. C. Two conformations of the integrin A-domain (I-domain): a pathway for activation? *Structure* **3**, 1333–1340 (1995).
27. Shimaoka, M. *et al.* Structures of the αI domain and its complex with ICAM-1 reveal a shape-shifting pathway for integrin regulation. *Cell* **112**, 99–111 (2003).
28. Scarborough, R. M. & Gretler, D. D. Platelet glycoprotein IIb-IIIa antagonists as prototypical integrin blockers: novel parenteral and potential oral antithrombotic agents. *J. Med. Chem.* **43**, 3453–3473 (2000).
29. Gottschalk, K. E. & Kessler, H. The structures of integrins and integrin-ligand complexes: implications for drug design and signal transduction. *Angew. Chem. Int. Edn Engl.* **41**, 3767–3774 (2002).
30. Egbertson, M. S. *et al.* Non-peptide GPIIb/IIIa inhibitors. 20. Centrally constrained thienothiophene alpha-sulfonamides are potent, long acting *in vivo* inhibitors of platelet aggregation. *J. Med. Chem.* **42**, 2409–2421 (1999).
31. Scarborough, R. M. *et al.* Design of potent and specific integrin antagonists. *J. Biol. Chem.* **268**, 1066–1073 (1993).
32. Mould, A. P. *et al.* Integrin activation involves a conformational change in the αI helix of the β subunit A-domain. *J. Biol. Chem.* **277**, 19800–19805 (2002).
33. Chen, J. F., Salas, A. & Springer, T. A. Bistable regulation of integrin adhesiveness by a bipolar metal ion cluster. *Nature Struct. Biol.* **10**, 995–1001 (2003).
34. Perutz, M. F. Mechanisms of cooperativity and allosteric regulation in proteins. *Q. Rev. Biophys.* **22**, 139–237 (1989).
35. Love, C. A. *et al.* The ligand-binding face of the semaphorins revealed by the high-resolution crystal structure of SEMA4D. *Nature Struct. Biol.* **10**, 843–848 (2003).
36. Bork, P., Doerks, T., Springer, T. A. & Snel, B. Domains in plexins: Links to integrins and transcription factors. *Trends Biochem. Sci.* **24**, 261–263 (1999).
37. Calvete, J. J., Henschen, A. & González-Rodríguez, J. Assignment of disulphide bonds in human platelet GPIIIa. A disulphide pattern for the β -subunits of the integrin family. *Biochem. J.* **274**, 63–71 (1991).
38. Beglova, N., Blacklow, S. C., Takagi, J. & Springer, T. A. Cysteine-rich module structure reveals a fulcrum for integrin rearrangement upon activation. *Nature Struct. Biol.* **9**, 282–287 (2002).
39. Newman, P. J., Derbes, R. S. & Aster, R. H. The human platelet alloantigens, PlA1 and PlA2, are associated with a leucine33/proline33 amino acid polymorphism in membrane glycoprotein IIIa, and are distinguishable by DNA typing. *J. Clin. Invest.* **83**, 1778–1781 (1989).
40. Watkins, N. A. *et al.* HPA-1a phenotype-genotype discrepancy reveals a naturally occurring Arg93Gln substitution in the platelet $\beta 3$ integrin that disrupts the HPA-1a epitope. *Blood* **99**, 1833–1839 (2002).
41. Kunicki, T. J. *et al.* The P1A alloantigen system is a sensitive indicator of the structural integrity of the amino-terminal domain of the human integrin $\beta 3$ subunit. *Blood Cells Mol. Dis.* **21**, 131–141 (1995).
42. Chen, J. F. *et al.* The relative influence of metal ion binding sites in the I-like domain and the interface with the hybrid domain on rolling and firm adhesion by integrin $\alpha 4\beta 7$. *J. Biol. Chem.* (in the press).
43. Luo, B.-H., Takagi, J. & Springer, T. A. Locking the $\beta 3$ integrin I-like domain into high and low affinity conformations with disulfides. *J. Biol. Chem.* **279**, 10215–10221 (2004).
44. Yang, W., Shimaoka, M., Chen, J. F. & Springer, T. A. Activation of integrin β subunit I-like domains by one-turn C-terminal α -helix deletions. *Proc. Natl Acad. Sci. USA* **101**, 2333–2338 (2004).
45. Takagi, J., Erickson, H. P. & Springer, T. A. C-terminal opening mimics “inside-out” activation of integrin $\alpha 5\beta 1$. *Nature Struct. Biol.* **8**, 412–416 (2001).
46. Esnouf, R. M. An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.* **15**, 132–138 (1997).
47. Merritt, E. A. & Murphy, M. E. P. Raster 3D version 2.0: a program for photorealistic graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).
48. Carson, M. Ribbons. *Methods Enzymol.* **277**, 493–505 (1997).
49. Puzon-McLaughlin, W., Kamata, T. & Takada, Y. Multiple discontinuous ligand-mimetic antibody binding sites define a ligand binding pocket in integrin $\alpha IIb\beta 3$. *J. Biol. Chem.* **275**, 7795–7802 (2000).
50. Tozer, E. C., Liddington, R. C., Sutcliffe, M. J., Smeeton, A. H. & Loftus, J. C. Ligand binding to integrin $\alpha IIb\beta 3$ is dependent on a MIDAS-like domain in the $\beta 3$ subunit. *J. Biol. Chem.* **271**, 21978–21984 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank colleagues in the Springer laboratory for supporting data and stimulating discussions, B. Kessler for tandem mass spectrometry, E. Yvonne Jones at Oxford for sema4D coordinates, members of the J.H.W. and M. Eck laboratories and the staff at APS and CHESS for assistance with crystallography, M. Gerstein and N. Echols (Yale University) for the morphing script used in producing movies, and Y. Cheng for help with comparing crystal structures and the electron microscopy map. Supported by NIH grants to T.A.S., J.H.W. and B.S.C.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to T.A.S. (springer@cbr.med.harvard.edu). Atomic coordinates and structural factors have been deposited with the Protein Data Bank with the accession codes 1TY3 (form A + cacodylate 1); 1TXV (form A + cacodylate 2); 1TY5 (form A + tirofiban); 1TY6 (form A + eptifibatide); 1TY7 (form A + L-739758); 1TYE (form B + cacodylate). Sequences of the 10E5 Fab have been deposited with GenBank.

Insights into assembly from structural analysis of bacteriophage PRD1

Nicola G. A. Abrescia^{1*}, Joseph J. B. Cockburn^{1,2*}, Jonathan M. Grimes¹, Geoffrey C. Sutton¹, Jonathan M. Diprose¹, Sarah J. Butcher³, Stephen D. Fuller¹, Carmen San Martín^{4†}, Roger M. Burnett⁴, David I. Stuart^{1,2}, Dennis H. Bamford³ & Jaana K. H. Bamford³

¹Division of Structural Biology, The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford OX3 7BN, UK

²Oxford Centre for Molecular Sciences, Central Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QT, UK

³Institute of Biotechnology and Faculty of Biosciences, Viikki Biocenter, University of Helsinki, P.O. Box 56, Viikinkaari 5, 00014, Finland

⁴The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104, USA

* These authors contributed equally to this work

† Present address: Biocomputing Unit, Centro Nacional de Biotecnología, Campus Universidad Autónoma, 28049 Madrid, Spain

The structure of the membrane-containing bacteriophage PRD1 has been determined by X-ray crystallography at about 4 Å resolution. Here we describe the structure and location of proteins P3, P16, P30 and P31. Different structural proteins seem to have specialist roles in controlling virus assembly. The linearly extended P30 appears to nucleate the formation of the icosahedral facets (composed of trimers of the major capsid protein, P3) and acts as a molecular tape-measure, defining the size of the virus and cementing the facets together. Pentamers of P31 form the vertex base, interlocking with subunits of P3 and interacting with the membrane protein P16. The architectural similarities with adenovirus and one of the largest known virus particles PBCV-1 support the notion that the mechanism of assembly of PRD1 is scaleable and applies across the major viral lineage formed by these viruses.

PRD1 is a double-stranded DNA bacteriophage and prototype member of the Tectiviridae. It infects Gram-negative bacteria, such as *Escherichia coli* and *Salmonella enterica*. The mature virion has a molecular mass of 66 MDa and comprises over 20 distinct protein species¹. Protein P3 (43 kDa) is the major component of the capsid, with 240 trimers arranged on an icosahedral lattice with pseudo-T = 25 triangulation. Pentamers of protein P31 (14 kDa) occupy the icosahedral vertices and associate with the trimeric P5 protein (34 kDa) and the receptor-binding protein P2 (64 kDa) to form flexible spikes². The minor capsid protein P30 (9 kDa) is required for virus assembly³. The 15-kilobase linear viral genome is packaged into preformed procapsids⁴ through a unique vertex^{5,6}, fuelled by the ATPase P9 (J.K.H.B., unpublished data). PRD1 is very different from the traditional DNA bacteriophages: instead of a tail it uses its internal membrane, acquired from the host during virus assembly^{4,7}, as an ejection device^{8,9}. In addition, there is no critical capsid expansion after DNA packaging¹⁰. However, PRD1 is remarkably similar to adenovirus, and several lines of evidence, including the structural homology between the major structural proteins, support the hypothesis that the two viruses have a common ancestor^{11–13}. Recently, the same structural characteristics have been found in the giant PBCV-1 (Phycodnaviridae family) and CIV viruses (Iridoviridae family)^{13–15}, indicating that PRD1 belongs to an extensive lineage, whose origin might predate the division of the bacterial, archaeal and eukaryotic domains of life¹³. Previous cryo-electron microscopic (cryo-EM) analysis of the PRD1 virion and X-ray analysis of isolated P3 illuminated aspects of PRD1 architecture^{10,11,16,17} but the analysis presented here, using crystals of the PRD1 Sus539 mutant^{1,18}, reveals in atomic detail the key structural proteins, lays bare the organization of the capsid shell and reveals the membrane. The structure of the membrane is described in the accompanying paper¹⁹. Here we focus on the architecture of the PRD1 capsid, which suggests a scaleable assembly pathway that might apply to some of the largest known viruses and reveals receptor-binding vertex connections to the viral membrane.

Sus539 mutant is described in Methods. Phase information was obtained by combining an X-ray structure of the major capsid protein^{11,20} with a cryo-EM reconstruction of the virus^{10,16}. Neglecting magnification errors, the deviation between Cα atoms derived from EM/X-ray hybrid imaging and the final X-ray structure is less than 2 Å within an icosahedral asymmetric unit. The resolution of the resultant electron density map is sufficient to resolve the β-strands in protein P3 (Fig. 1a). In addition, data from crystals of selenomethionated (SeMet) virus provide a SeMet difference Fourier map. Using these two electron density maps we have constructed atomic models for the minor capsid protein P30, the vertex protein P31, the membrane protein P16 and previously undescribed portions of the major capsid protein (see Methods). The analysis is based on 60-fold averaged electron density in which proteins not obeying the icosahedral symmetry are smeared out and are essentially invisible. PRD1 measures about 640 Å between opposite faces and 700 Å between opposite vertices. The Sus539 mutant lacks P2, the receptor-binding protein, whereas the trimeric P5 is not seen in the averaged electron density.

The 240 trimers of protein P3 dominate the PRD1 capsid (Fig. 1b). Each P3 subunit is composed of two ‘jellyroll’ domains^{11,20}, giving the trimer a pseudo-hexagonal shape. The icosahedral asymmetric unit of the virus is composed of four P3 trimers (Fig. 1b) arranged in a uniform orientation within the roughly planar facet (root-mean-square (r.m.s.) deviation from planarity 6.9 Å), to give the pseudo-T = 25 lattice characteristic of PRD1 and adenovirus. The cores of the 12 independent copies of P3 are essentially identical to each other and to the single crystal structures of isolated P3 (ref. 11) (r.m.s. deviation in Cα atoms, after superposition, is about 1 Å). There are, however, different conformations for the short amino-terminal and carboxy-terminal extensions, depending on the location of the subunit in the surface lattice of the virus, leading to small differences in the conformation of some loops (Fig. 1c).

We find that the capsid contains 60 copies of the small protein P30 (83 residues). This protein is essential for the assembly of complete particles³. P30 is remarkably extended, with essentially no secondary structure and dimensions of about 150 × 60 × 10 Å³

Overall architecture of the virus and structural proteins

The structure determination at about 4 Å resolution for the PRD1

(Fig. 1d). The electron density of P30 is continuous, and detailed structural interpretation was assisted by the SeMet difference Fourier map, which pinpointed the methionine residues at positions 24, 45 and 69 (see Fig. 1d). The molecule is rich in proline residues (13%) and analysis of its amino acid sequence with PONDR²¹ indicates that in the absence of other viral components at least the N-terminal half of the subunit would be disordered. The 60 copies of P30 completely festoon the virus, being suspended about 12 Å above the viral membrane by numerous interactions with P3 (Fig. 1d). N-terminal hooks (the only deviation from a fully extended conformation) lock two P30 subunits into the dimeric building block of the network (dimer interactions are restricted to

about the first 30 residues of the subunit). Dimer axes coincide with icosahedral two-fold axes and the protein follows closely the outline of the P3 facet (P30/P3 interactions are described below).

Pentamers of protein P31 plug the 60-Å diameter holes in the P3/P30 shell at the five-fold vertices. P31 forms a compact eight-stranded β -barrel with jellyroll topology (Fig. 1e), elaborated only by 13-residue and 8-residue extensions at the N and C termini respectively. Residues 5–118 of the 126-amino-acid subunit have been modelled with the aid of the SeMet difference Fourier map. P31 resembles half of a P3 subunit and can be superposed almost equally well on either of the P3 jellyrolls (82 and 83 C α atoms superposed for domains 1 and 2, respectively, in each case with an

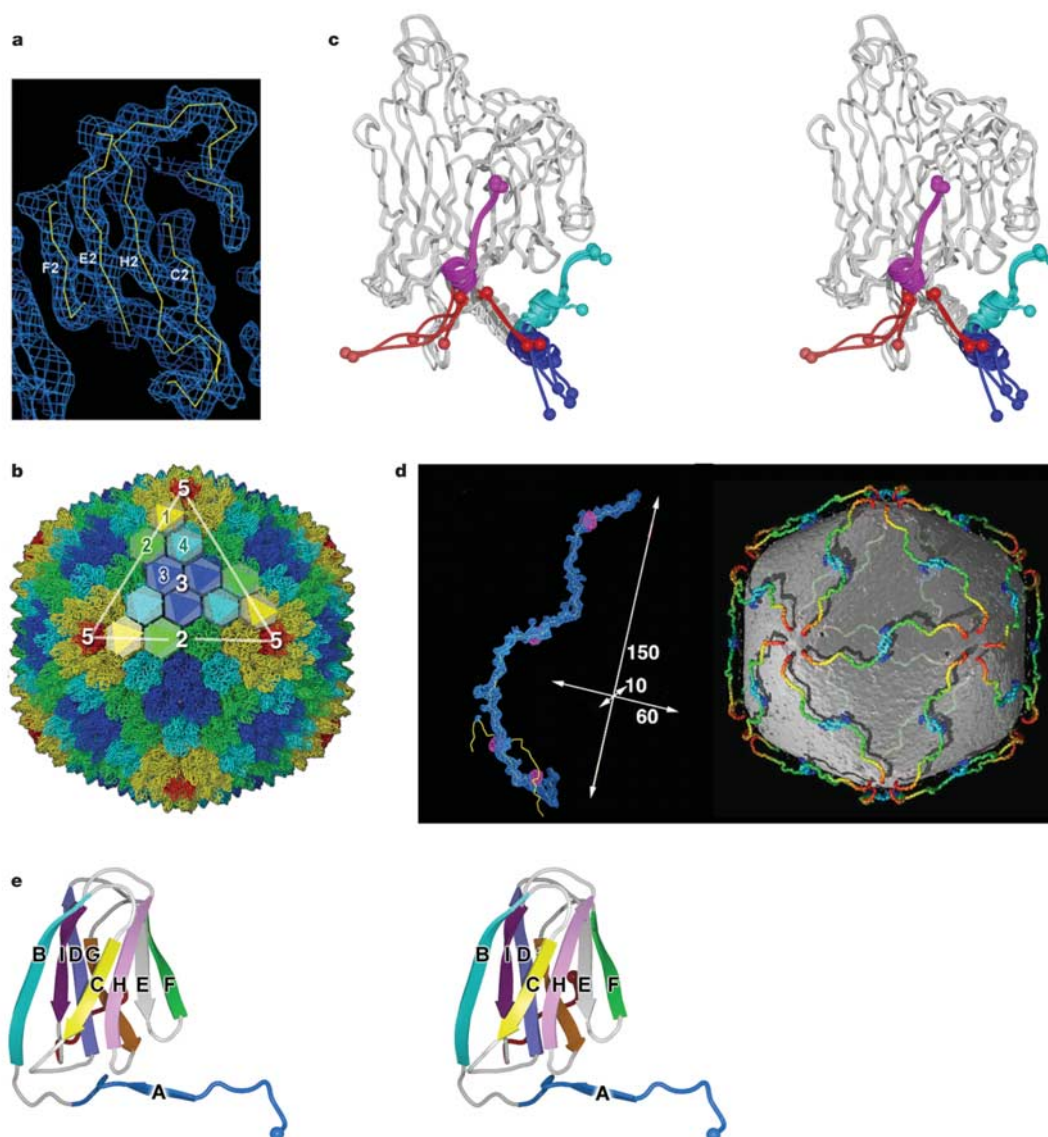


Figure 1 Architecture and structural components of bacteriophage PRD1. **a**, Electron density of cell 2 contoured at 1.5 σ . The strands of the P3 second jellyroll are clearly resolved. **b**, The building blocks of a facet (the triangular area defined by white lines). The icosahedral asymmetric unit contains 12 P3 subunits arranged as four trimers, with orientations represented by triangles labelled 1 to 4 and coloured yellow, green, blue and cyan respectively (the underlying grey hexagons show the trimer morphology; numbers and symbols follow the adenovirus convention³⁹). P3 trimers outside the marked facet are coloured similarly and shown as coil, as is P31, drawn in red. **c**, Stereo C α traces of the superposed 12 unique copies of P3, showing switching at the N (blue and cyan) and C (red and magenta) termini. **d**, Protein P30. Left: royal blue, 2F_o - F_c electron density

contoured at 1.2 σ ; magenta, SeMet difference density contoured at 8.5 σ (the peaks, from the top, correspond to SeMet residues 69, 45 and 24). Part of a two-fold related P30 is shown in yellow. Dimensions are in ångströms. Right: 60 copies of P30 (coloured blue through green to red from the N terminus to the C terminus) wrap around the electron density of the membrane. The small holes in the membrane close to the five-fold axes are the transmembrane helices of P16. **e**, Stereo view of P31. Jellyroll strands are labelled B to G; the β -strand at the N terminus is labelled A. N-terminal and C-terminal segments are coloured blue and red respectively. β -strand residues: B, 22–31; C, 34–41; D, 48–55; E, 60–66; F, 69–76; G, 79–87; H, 90–97; I, 101–109.

r.m.s. deviation of 2.9 Å as determined with program SHP²²), although there is no sequence similarity between them. Finally, beneath each P31 pentamer lie five copies of protein P16. Residues 7–56 and 94–116 are visible in the electron density, with residues 7–28 forming a transmembrane helix. The assignment of P16 is described in Methods.

Conformational switching in P3 underlies capsid assembly

In closed assemblies, geometric constraints prevent an exactly symmetric arrangement of more than 60 subunits. The switching of subunit conformation during the assembly process provides a mechanism for building more complex systems (see, for example, refs 23 and 24), and this was suggested for PRD1 (ref. 17). The 720 subunits of P3 present in each PRD1 particle occupy 12 distinct

local environments within each facet (Fig. 2), which define the conformations of the N-terminal and C-terminal extensions. Whereas in the intact virus these extensions are generally well ordered, in the crystal structure of isolated P3 both termini were disordered (about 13 and 11 residues respectively)^{11,20} implying that there is little energetic penalty in switching between conformers. The N termini of P3 fall into two distinct conformations (denoted NI and NII) that converge at residue 18 (between 9 and 12 Å above the outer leaflet of the membrane). Residues 18–34 form an α-helix that runs away from the membrane. In one set of conformations (NII, shown in dark blue in Figs 1c and 2) this helix is longer, beginning at residue 8, and is preceded by a short extended section of chain that contacts the membrane (Fig. 2). Three of the first six amino acids are glutamine residues, which probably attach to the

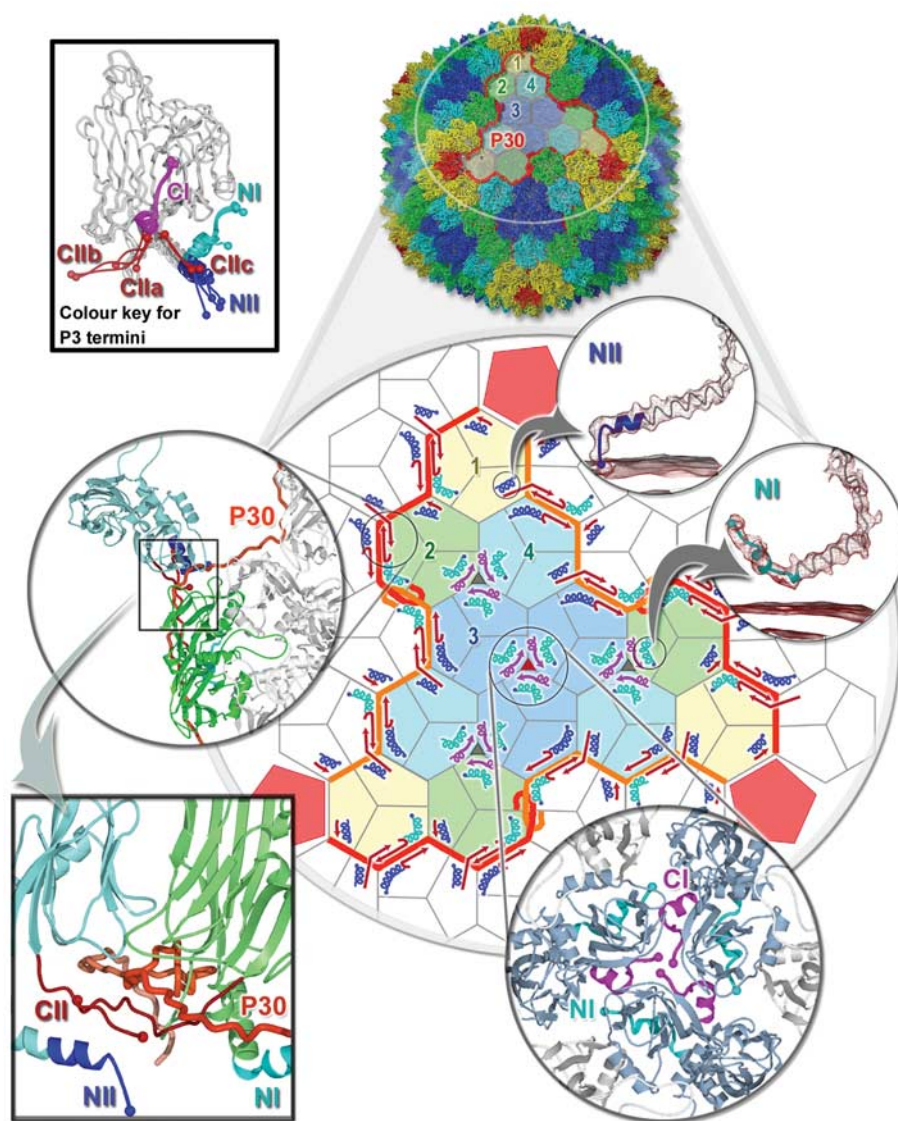


Figure 2 Switching of the P3 termini. Top left, key summarizing the conformations. Top centre, overview of the particle. The white oval marks the portion shown below. Centre, schematic diagram of a facet with neighbouring P3 trimers (white). Trimers are shown as colour-coded hexagons. The P31 pentamer is shown as a bright pink pentagon. Two-fold related P30 subunits are drawn in red and orange. Three-fold and pseudo-three-fold axes are depicted as small triangles, red and grey respectively. For each P3 subunit, the N terminus is shown as a straight or kinked helix and the C terminus as a small arrow,

coloured according to the key. Type CIIa termini (the short hooks in the main diagram) form a two-stranded β-structure with P30. Type CIIb termini are further ordered, taking part in a three-stranded β-sheet; the other two strands are contributed by P30 and by the C termini of P3 subunits from the neighbouring facet (type CIIc, shown as doglegs in the main diagram). The close-ups linked by straight zoom-lines are viewed to match the diagram, and grey arrows indicate an orthogonal view. Close-ups on the right include electron density (contoured at 0.6σ) for the N termini and membrane outer leaflet.

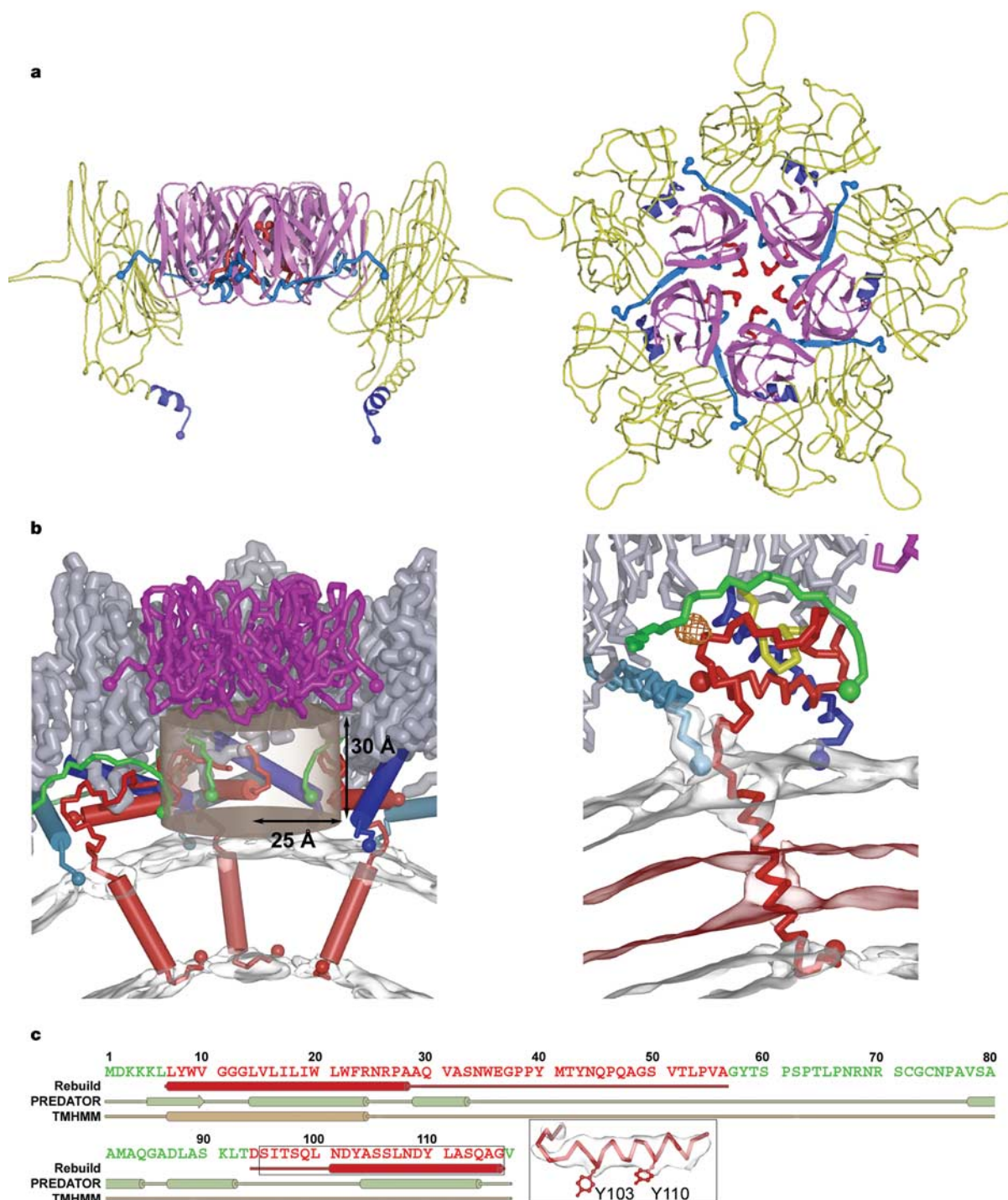


Figure 3 Vertex organization. **a**, Protein P31 and peripentonal P3 subunits. P31 is shown in lilac. The P31 N and C termini are shown as royal blue and red coils respectively. Peripentonal P3 subunits are drawn as yellow coils (residues involved in conformational switching at the N terminus, which contact the membrane, are dark blue). Left, view orthogonal to the five-fold axis. Right, view along the five-fold axis, showing the interdigitation of the N terminus of P31 (royal blue) at the interface between peripentonal P3 monomers. **b**, P16 and its interactions. Left, five copies of P16 (red) are arranged around each vertex, between the peripentonal P3 trimers (grey), of which three are shown. Each copy of P16 contributes disordered residues to the 25 Å × 30 Å cylindrical cavity under the vertex base protein P31 (magenta). The first six residues of P16 beneath the membrane are disordered but are shown for completeness. Right, each copy of P16 interacts with two subunits of the adjacent peripentonal P3 trimer clockwise around the five-fold axis and the P30 C-terminal region (green). The N terminus of the peripentonal

P3 subunit furthest from the vertex (light blue) contacts P16 at the membrane. The loop connecting the two jellyrolls of the P3 subunit nearest the vertex (yellow) is clipped between P16 residues 43–54 and the P16 C-terminal helix, which is flanked on the other side by the corresponding P3 N terminus (dark blue). Protein termini are represented by balls. The outer and inner leaflet headgroup regions of the membrane are contoured in grey at 0.6σ and 0.3σ, respectively, with the transmembrane helix density contoured in red at −0.3σ. The selenomethionine peak for P16 residue 41 (which identified the protein) is contoured in orange at 8.5σ (see Methods). **c**, Secondary structure elements of our P16 model (red), secondary structure prediction (PREDATOR⁴⁹, green) and transmembrane helix prediction (TMHMM⁴⁷, brown), aligned with the amino acid sequence (amino acids visible in the electron density are in red). The electron density for the P16 C-terminal helix, contoured at 1.5σ, shows two large side chains consistent with the tyrosine residues at positions 103 and 110 (see Methods).

headgroups of the outer leaflet of the membrane. These membrane-anchoring P3 subunits are found along the edge of the icosahedral facets and might help to stabilize the fold in the membrane that occurs at this intersection¹⁹. In the other N-terminal conformation (NI, shown in cyan in Figs 1c and 2), the residues preceding 18 form a short α -helix twisted away from the membrane (Fig. 2) to interact with another subunit of the P3 trimer and, for subunits within the body of the facet, with the C termini (residues 385–390) of P3 subunits from neighbouring trimers. This conformational switching allows the P3 N termini to stabilize either the viral membrane or interactions with neighbouring P3 molecules. The switching of the P3 C termini is central to the assembly process, because these structures mediate P3 trimer–trimer interactions. As with the N-terminal switching, the four subunits within the body of the icosahedral face have essentially the same structure (denoted CI in Figs 1c and 2). In this conformation, after a short helix (residues 383–389) the final six residues extend to form a trimeric interaction with termini from neighbouring trimers, locking a set of three trimers together (Fig. 2). The electron density and the presence of acidic residues suggest that this structure might be further stabilized by a divalent cation lying on the local inter-trimer three-fold axis. In conformation CII (brick red on Figs 1c and 2), adopted by six subunits, the helix is replaced by an extended chain. These extensions (which constitute three subsets, CIIa–c; see Fig. 2) form β -structures with P30, which threads between the P3 trimers of neighbouring facets. P30 is clamped by P3 in four β -sheets along the edge of the facet (Fig. 2), stabilizing the facet interface. The final two P3 C-terminal structures (short red arrows in Fig. 2, found in the peripentonal P3 trimers and subunits abutting them) are disordered, eliminating C-terminal interactions between the peripentonal trimers and the rest of the structure.

Interactions at the vertex

Five subunits of P31 pack with their jellyrolls vertical to form the pentameric base of the vertex spike (Fig. 3a). The pentamer is stabilized by residues 9–14 (strand A in Fig. 1e) clipping onto strand B of the adjacent P31 β -barrel and by side-to-side interactions between the BIDG sheet of one subunit and strands G and F of its

neighbour. A series of interactions engage the P31 pentamer with the P3 subunits that surround it. The nine N-terminal residues of P31 interdigitate between the B1–C1 and B2–C2 β -strands of five-fold related P3 subunits, and the ACHEF face of P31 nestles between the two jellyrolls of a peripentonal P3 subunit (Fig. 3a).

There is a symmetry mismatch between the pentameric P31 and the trimeric spike protein P5 (a similar symmetry mismatch is found in adenovirus²⁵). P31 and the N-terminal domain of P5 share 38% sequence identity^{26,27}, which might allow a mixed pentameric association of P31 with the N-terminal domains of P5 (ref. 28). However, there is little evidence for this in the electron density maps: P31 is generally well defined, although the attenuation of the SeMet difference Fourier peak for residue Met9 would be consistent with up to two of the five subunits being P5. The C termini of P31 lie close to the five-fold axes, in line with direct P5–P31 interactions at these axes^{17,27}.

Below P31, in a position where unidentified density was seen in cryo-EM analysis¹⁷, lies the 117-residue integral membrane protein P16 (ref. 29) (Fig. 3b, c), anchored into the membrane through a single transmembrane helix (residues 7–28). Residues 1–6 are disordered beneath the membrane, with the triple lysine sequence 3–5 presumably stabilizing the negative charge of the viral DNA. There is no electron density for residues 57–93, indicating that this portion (denoted loop_{57–93}) does not obey icosahedral symmetry. The trajectory of the polypeptide chain entering and leaving loop_{57–93} indicates that it resides in a cylindrical cavity under the vertex (Fig. 3b), which it largely fills (see Supplementary Methods). P16 engages in extensive interactions with the P3 trimers adjacent to the vertex and the C-terminal region of the cementing protein P30 (Fig. 3b). The vertex complexes and peripentonal P3 trimers are labile in PRD1 virions lacking P16 (ref. 29), whereas removal of the phage membrane from PRD1 particles lacking DNA, by treatment with sodium dodecyl sulphate, removes the vertex complex and peripentonal P3 trimers³⁰. Thus P16 acts as a second cementing protein stabilizing the vertices, a function analogous to that of adenovirus polypeptide VI (ref. 31). The observed lability of the vertex might facilitate the formation of a pore in the capsid for egress of the PRD1 DNA injection tube³². The disordered loop_{57–93} probably forms non-icosahedrally symmetric interactions with the vertex base, presumably explaining arms of electron density seen connecting the P16 region and capsid vertex in the cryo-EM analysis¹⁷. The breakdown in icosahedral symmetry might reflect the symmetry mismatch between the pentameric vertex base and the trimeric spike shaft. Perhaps P16 acts as the component that transmits the receptor-binding signal to the membrane, priming the metastable system for DNA ejection². In addition to P16, 11 further protein species have been reported to be associated with the viral membrane^{1,4,32–35}; in particular, two integral membrane proteins and two capsid-associated proteins are located at a unique vertex responsible for DNA packaging^{5,6}. We see no evidence of these proteins in the averaged electron density maps and unaveraged electron density maps show no features to suggest a preferred orientation of the special vertices within the crystal lattice. Visualization of the remaining membrane proteins is therefore probably beyond the reach of the current data.

Table 1 Data collection statistics

Location	Reflections	Resolution (Å)	$I/\sigma(I)^*$	$R_{\text{merge}}^\dagger$	Completeness (%)
Cell 1	3829761‡	100–4.2	1.8	0.33	45
	2383561§	5.05–4.98	1.1	0.63	42
		4.27–4.20	0.9	0.65	7
Cell 2	1088596‡	95.3–4.2	2.6	0.23	16
	899751§	5.05–4.98	1.3	0.45	8
		4.48–4.43	1.0	0.50	3
SeMet	5180365‡	67.4–3.7	1.49	0.42	43
	3345164§	4.27–4.17	0.96	0.59	32
		3.76–3.70	0.70	0.67	3

* σ values were adjusted to reflect the observed discrepancies between multiply measured data.
† $R_{\text{merge}} = \sum_i \sum_j |I_{ij} - \langle I_i \rangle| / \sum_i \sum_j I_{ij}$, where I_{ij} are the unique reflection intensities, I_i are the intensities of symmetry related reflections and $\langle I_i \rangle$ is the mean intensity.
‡ Measured reflections.
§ Unique reflections.

Table 2 Phase refinement statistics for the merged native–SeMet cell 1 and cell 2 data sets

Resolution (Å)	Merged cell 1			Cell 2		
	Completeness (%)	Averaging R -factor*	Correlation coefficient†	Completeness (%)	Averaging R -factor*	Correlation coefficient†
100–4.2	57	0.33	0.55	16	0.18	0.90
5.01–4.97	54	0.36	0.25	7	0.21	0.80
4.23–4.21	17	0.34	0.26	1	0.18	0.77

* R -factor = $\sum_h |F_o| - |F_c| / \sum_h |F_o|$, where h are the unique reflection indices, F_o are the observed structure factors and F_c are the structure factors obtained from inversion of the solvent-flattened map (SHELLSCALE; D.I.S., unpublished program).
† Correlation coefficient = $\sum_h (|F_o| - |F_c|)(|F_o| - |F_c|) / \sum_h (|F_o| - |F_c|)^2 - ((|F_o| - |F_c|)^2)^{1/2}$, where h are the unique reflection indices, F_o are the observed structure factors and F_c are the structure factors obtained from inversion of the solvent-flattened map (SHELLSCALE; D.I.S., unpublished program).

P30 and a model for viral assembly

P30 cements P3 along the edge of the facet (Fig. 2), with the C-terminal tail of P30 reaching the penton where it also interlocks with protein P16 (at residues 70–79 of P30). Large viruses frequently use a scaffolding protein to drive assembly and control the capsid size³⁶, for example gp8 of bacteriophage P22 and gp7 of bacteriophage ϕ 29, both of which are elongated proteins with a substantial α -helical content^{36,37}. P30 of PRD1 is even more elongated, being almost completely extended, allowing the dimer to span the entire edge of an icosahedral facet (about 300 Å). In effect P30 acts as a molecular tape-measure, a function attributed to protein YscP in the assembly of the bacterial injectisome, and also responsible for the determination of phage tail length (ref. 38 and references therein). The linear nature of the 83-residue P30 would allow a virus of eightfold the mass of PRD1 to be assembled with a minimal 'P30' of less than 200 residues, and very large viruses (ref. 14, and S. D. Benson, J. K. H. Bamford, D. H. Bamford and R. M. Burnett, unpublished data) could be assembled with a tape-measure protein of only 300 amino acids. The efficient and accurate assembly of such complex structures with many thousands of subunits presents numerous challenges, but the limited conformational switching seen in PRD1 P3 suggests a plausible pathway in which assembly is governed by extensive β -sheet interactions between P3 and P30. The use of preformed P3 trimers and P30 dimers probably provides early quality-control steps in the assembly pathway. A plausible assembly model could involve the initial attachment of two copies of P3 trimer 2 (defined in Fig. 2) around the P30 dimer axis, followed by the attachment of trimers 3 and 1 (probably with fairly independent kinetics) to form the binding face to which trimer 4 then binds. Such an assembly of eight trimers of P3 and a dimer of P30 would represent 1/30 of the virus, and these preformed units might then clip together to form the complete virus shell. P31, P16 and perhaps other proteins are likely to be involved in the final stages of the assembly process. Particle formation is also aided by two non-structural assembly factors, P10 and P17 (ref. 32), with a large amount of P10 being associated with the viral membrane during capsid assembly³. We suggest that P10 is involved in the formation of correctly sized membrane vesicles, which the assembling procapsid selects through a size-exclusion mechanism to form an empty virion into which the genome is packaged through a single special vertex^{6,32}.

Generic features of assembly of a major virus lineage

The structure of the major capsid protein of the massive PBCV-1 (ref. 15) validated the proposal¹³ that the Phycodnaviridae and Iridoviridae belong to the PRD1/adenovirus structure-based lineage, and analysis (S. D. Benson, J. K. H. Bamford, D. H. Bamford and R. M. Burnett, unpublished data) suggests that the 'double-barrelled' major capsid protein is present in other virus families including the Asfarviridae, the Ascoviridae and the very large Mimivirus. However, the set of core properties that defines the PRD1/adenovirus lineage might extend beyond the architecture of the major structural protein. The PRD1 vertex protein, P31, resembles a single jellyroll domain of P3, and presumably these proteins originated from a common ancestor constructed from a single jellyroll. The divergence of P3 and P31 probably occurred before the separation of adenovirus from PRD1, so the adenovirus penton base might possess a much elaborated jellyroll (molecular mass 63 kDa, in comparison with 13 kDa for P31), echoing the relationship between P3 of PRD1 and the adenovirus hexon¹¹. The pseudo-hexagonal arrangement of structurally homologous trimeric molecules into triangular facets is another core property of the PRD1/adenovirus lineage, which probably reflects underlying similarities in the assembly mechanism^{13–15}. Thus, adenovirus might possess a homologue of P30. Although postulated to have an analogous function³¹, polypeptide IIIa is larger than expected for

a P30 homologue and is a less attractive candidate than polypeptide VIII, which is proline-rich, is almost exactly the expected length (110 residues) and possesses an N-terminal sequence that PONDR²¹ predicts to be disordered. The larger capsids such as PBCV-1 are, like PRD1 and adenovirus, assembled principally out of triangular plates composed of trimers of a single protein arranged with $p3$ symmetry³⁹, with neighbouring plates related by icosahedral two-fold axes, suggesting that all these viruses might use a tape-measure mechanism to control assembly. Further structural analyses are required to test this hypothesis.

Discussion

Bacteriophage PRD1 adopts a simple and effective strategy for protein shell assembly. The bulk of the capsid is cemented together by protein P30, and the vertex is stabilized by the integral membrane protein P16. The latter seems to act as a metastable sensor, linking the symmetry-mismatched vertex base and receptor-binding spike shaft with the membrane and providing a mechanism for the controlled extrusion of a membrane–protein–DNA tube to initiate host entry. Particle assembly seems to be regulated by P30, which acts as a molecular tape-measure, modulating a two-state conformational switch in the major capsid protein (P3) to control both the assembly of an initial P30–P3 core structure and the proliferation of P3–P3 contacts. This allows 60 subunits of the small, essentially linear, P30 protein to control the assembly of 720 P3 subunits. This model can be easily scaled to very large systems, and circumstantial evidence indicates that it might underpin some of the most complex self-assembling systems found in biology. □

Methods

Crystallographic data collection and processing

The growth, purification and crystallization of PRD1 Sus539 mutant has been described^{14,18}. For SeMet labelling, the virus was grown on M9 glucose–thiamine medium⁴⁰ and 60 mg ml^{−1} L-selenomethionine (Calbiochem) added 20 min after infection. Diffraction data were collected at ~21 °C on beamlines ID13 and ID14 of the European Synchrotron Radiation Facility (ESRF), Grenoble^{14,18}. Data were analysed with DENZO⁴¹ and the resolution limit was determined with TRIM_DENZO (D.I.S., unpublished program). Data were scaled with a special version of SCALEPACK⁴¹. For the native data and SeMet data, reflections with fractional partialities of 0.7 or more and 0.5, respectively, were scaled to full intensity and incorporated into the data set as fully recorded reflections (program POST: D.I.S. and J.M.D., unpublished program). The data from native crystals fall into two sets with the same spacegroup $P2_12_12_1$ but different unit cell dimensions¹⁸. The SeMet-substituted crystals were isomorphous with native cell 1. The crystallographic asymmetric unit contains an entire virus particle. Data statistics are summarized in Table 1 and Supplementary Figure. Intensities were converted to structure factor amplitudes with TRUNCATE⁴².

Phasing and phase refinement

Initial phases were obtained by molecular replacement with the use of a quasi-atomic model of the viral capsid (PDB entry 1HB7) derived from fitting a 14 Å cryo-EM reconstruction¹⁶. The position and orientation of this model were refined against the experimental data using X-PLOR 3.1 (ref. 43). This was performed separately for data from the cell 1 and cell 2 crystal forms. An initial estimate of the particle orientation was obtained from the self-rotation function calculated from the cell 1 data set. Patterson correlation refinement⁴⁴ was performed against 60–15-Å resolution data to optimize the orientation (final correlation coefficient (cc) of 0.021 for cell 1 and 0.027 for cell 2). The position and orientation of the particle were then refined at 30 Å with the E1E1 target function⁴⁵ (cc = 0.196 for cell 1, 0.195 for cell 2), giving an initial set of non-crystallographic symmetry (NCS) operators for the two cell types. P3 trimers inside the icosahedral asymmetric unit were refined at 6 Å resolution and NCS operators were re-refined at 8 Å resolution⁴⁵ (cc = 0.375 for cell 1, 0.546 for cell 2). Further refinement of the trimer positions and the NCS operators with data to 5 Å gave final correlation coefficients of 0.230 and 0.423 for cell 1 and cell 2, respectively. A $2F_o - F_c$ map was calculated for each cell type (to 4.2 Å resolution) with the corresponding model phases, which were then refined by iterative solvent flattening and 60-fold averaging; F_c was calculated from the starting model or from Fourier transformation of the modified map (see Table 2) (program GAP; D.I.S. and J.M.G., unpublished program, and CCP4 programs⁴²). For missing reflections, including those at low resolution, appropriately weighted F_c values were used.

Calculation of the SeMet difference map

The SeMet and native cell 1 data were scaled in resolution shells with the use of SHELLSCALE (D.I.S., unpublished program). Difference Fourier maps were calculated using cell 1 phases and difference attenuation weighting to down-weight terms with large experimental errors⁴⁵. The maps were 60-fold averaged. The mean (μ) and standard

deviation (σ) of the electron density was evaluated in hollow shells (outer and inner radii 30 and 15 Å, respectively) around each methionine residue in the P3, P30, P31 and P16 models, as well as the height of the difference peak ($\rho_{\max}$) and the signal-to-noise ratio ($\rho_{\max} - \mu/\sigma$). Varying the partiality threshold and resolution showed that the difference map calculated to 4.2 Å with partiality threshold 0.5 gave the highest mean signal-to-noise ratio (9.4).

Model building

The P3 N- and C-terminal regions were built into the averaged map of cell 2 by using O⁴⁶. Major movements of the loop regions of the different P3 subunits were fitted manually to the electron density. The SeMet difference Fourier showed five other strong peaks in the icosahedral asymmetric unit in addition to those attributable to P3. The location of three of these and the quality of the native map led to the unequivocal identification and tracing of the minor capsid protein P30 (residues 1–83). The electron density close to the five-fold axes was also interpreted. A jellyroll fold was apparent and a preliminary model was manually fitted as a rigid body, followed by retracing to match the observed map. The unique SeMet peak at position 9 identified this protein as the pentameric vertex protein P31 (residues 5–118).

The electron density maps revealed two apparently disconnected polypeptide fragments located between the peri-pentonal P3 trimers (Fig. 3b). Fragment I is a roughly 50-residue peptide containing a transmembrane helix about 20 residues long and a methionine residue (identified in the SeMet difference Fourier) about 11 residues from the point of contact with the membrane. Fragment II consists of about 23 amino acids, terminating in a roughly 17-residue α -helix. Analysing the relative positions of methionine residues and predicted transmembrane helices (program TMHMM⁴⁷) in all the candidate sequences from the virus showed that fragment I belonged to the N-terminal region of the 117-residue integral membrane protein P16, predicted to possess an 18-residue transmembrane helix between residues 7 and 24. Placing an α -helix into the transmembrane electron density (Fig. 3b, right) and positioning the P16 residue Met 41 in the vicinity of the SeMet peak enabled the construction of a C α trace for P16 residues 7–56. To give the best fit to the electron density, the transmembrane helix was extended from residue 24 to residue 28. Fragment II was assigned as the P16 C terminus on the basis of protein packing density considerations in the cavity under the vertex (see Supplementary Methods) and the alignment of tyrosines at positions 103 and 110 (the largest amino acids in the C-terminal sequence) to two particularly pronounced bumps seven residues apart (Fig. 3c). All structures were brought into conformity with expected backbone conformations and side chains were built with program CALPHA⁴⁸.

Received 9 March; accepted 21 September 2004; doi:10.1038/nature03056.

1. Bamford, J. K. *et al.* Diffraction quality crystals of PRD1, a 66-MDa dsDNA virus with an internal membrane. *J. Struct. Biol.* **139**, 103–112 (2002).
2. Rydman, P. S. *et al.* Bacteriophage PRD1 contains a labile receptor-binding structure at each vertex. *J. Mol. Biol.* **291**, 575–587 (1999).
3. Rydman, P. S., Bamford, J. K. & Bamford, D. H. A minor capsid protein P30 is essential for bacteriophage PRD1 capsid assembly. *J. Mol. Biol.* **313**, 785–795 (2001).
4. Mindich, L., Bamford, D., McGraw, T. & Mackenzie, G. Assembly of bacteriophage PRD1: particle formation with wild-type and mutant viruses. *J. Virol.* **44**, 1021–1030 (1982).
5. Stromsten, N. J., Bamford, D. H. & Bamford, J. K. The unique vertex of bacterial virus PRD1 is connected to the viral internal membrane. *J. Virol.* **77**, 6314–6321 (2003).
6. Gowen, B., Bamford, J. K., Bamford, D. H. & Fuller, S. D. The tailless icosahedral membrane virus PRD1 localizes the proteins involved in genome packaging and injection at a unique vertex. *J. Virol.* **77**, 7863–7871 (2003).
7. Bamford, D. H., Caldentey, J. & Bamford, J. K. Bacteriophage PRD1: a broad host range dsDNA tectivirus with an internal membrane. *Adv. Virus Res.* **45**, 281–319 (1995).
8. Grah, A. M., Daugelavicius, R. & Bamford, D. H. Sequential model of phage PRD1 DNA delivery: active involvement of the viral membrane. *Mol. Microbiol.* **46**, 1199–1209 (2002).
9. Grah, A. M., Daugelavicius, R. & Bamford, D. H. The small viral membrane-associated protein P32 is involved in bacteriophage PRD1 DNA entry. *J. Virol.* **76**, 4866–4872 (2002).
10. Butcher, S. J., Bamford, D. H. & Fuller, S. D. DNA packaging orders the membrane of bacteriophage PRD1. *EMBO J.* **14**, 6078–6086 (1995).
11. Benson, S. D., Bamford, J. K., Bamford, D. H. & Burnett, R. M. Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. *Cell* **98**, 825–833 (1999).
12. Belnap, D. M. & Steven, A. C. 'Deja vu all over again': the similar structures of bacteriophage PRD1 and adenovirus. *Trends Microbiol.* **8**, 91–93 (2000).
13. Bamford, D. H., Burnett, R. M. & Stuart, D. I. Evolution of viral structure. *Theor. Popul. Biol.* **61**, 461–470 (2002).
14. Yan, X. *et al.* Structure and assembly of large lipid-containing dsDNA viruses. *Nature Struct. Biol.* **7**, 101–103 (2000).
15. Nandhagopal, N. *et al.* The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. *Proc. Natl Acad. Sci. USA* **99**, 14758–14763 (2002).
16. San Martín, C. *et al.* Combined EM/X-ray imaging yields a quasi-atomic model of the adenovirus-related bacteriophage PRD1 and shows key capsid and membrane interactions. *Structure* **9**, 917–930 (2001).
17. San Martín, C. *et al.* Minor proteins, mobile arms and membrane-capsid interactions in the bacteriophage PRD1 capsid. *Nature Struct. Biol.* **9**, 756–763 (2002).
18. Cockburn, J. J., Bamford, J. K., Grimes, J. M., Bamford, D. H. & Stuart, D. I. Crystallization of the membrane-containing bacteriophage PRD1 in quartz capillaries by vapour diffusion. *Acta Crystallogr. D* **59**, 538–540 (2003).
19. Cockburn, J. J. *et al.* Membrane structure and interactions with protein and DNA in bacteriophage PRD1. *Nature* doi:10.1038/nature03053 (this issue).
20. Benson, S. D., Bamford, J. K., Bamford, D. H. & Burnett, R. M. The X-ray crystal structure of P3, the

- major coat protein of the lipid-containing bacteriophage PRD1, at 1.65 Å resolution. *Acta Crystallogr. D* **58**, 39–59 (2002).
21. Li, X., Romero, P., Rani, M., Dunker, A. K. & Obradovic, Z. Predicting protein disorder for N-, C-, and internal regions. *Genome Informatics* **10**, 30–40 (1999).
 22. Stuart, D. I., Levine, M., Muirhead, H. & Stammers, D. K. Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 Å. *J. Mol. Biol.* **134**, 109–142 (1979).
 23. Liddington, R. C. *et al.* Structure of simian virus 40 at 3.8 Å resolution. *Nature* **354**, 278–284 (1991).
 24. Grimes, J. M. *et al.* The atomic structure of the bluetongue virus core. *Nature* **395**, 470–478 (1998).
 25. Chiu, W., Burnett, R. M. & Garcea, R. L. in *Structural Biology of Viruses* (eds Chiu, W., Burnett, R. M. & Garcea, R. L.) 209–238 (Oxford Univ. Press, New York, 1997).
 26. Caldentey, J., Tuma, R. & Bamford, D. H. Assembly of bacteriophage PRD1 spike complex: role of the multidomain protein P5. *Biochemistry* **39**, 10566–10573 (2000).
 27. Sokolova, A. *et al.* Solution structure of bacteriophage PRD1 vertex complex. *J. Biol. Chem.* **276**, 46187–46195 (2001).
 28. Xu, L., Benson, S. D., Butcher, S. J., Bamford, D. H. & Burnett, R. M. The receptor binding protein P2 of PRD1, a virus targeting antibiotic-resistant bacteria, has a novel fold suggesting multiple functions. *Structure* **11**, 309–322 (2003).
 29. Jaatinen, S., Viitanen, S., Bamford, D. H. & Bamford, J. K. The integral membrane protein P16 of bacteriophage PRD1 stabilizes the adsorption vertex structure. *J. Virol.* **78**, 9790–9797 (2004).
 30. Luo, C., Butcher, S. & Bamford, D. H. Isolation of a phospholipid-free protein shell of bacteriophage PRD1, an *Escherichia coli* virus with an internal membrane. *Virology* **194**, 564–569 (1993).
 31. Stewart, P. L., Fuller, S. D. & Burnett, R. M. Difference imaging of adenovirus: bridging the resolution gap between X-ray crystallography and electron microscopy. *EMBO J.* **12**, 2589–2599 (1993).
 32. Bamford, D. & Mindich, L. Structure of the lipid-containing bacteriophage PRD1: disruption of wild-type and nonsense mutant phage particles with guanidine hydrochloride. *J. Virol.* **44**, 1031–1038 (1982).
 33. Bamford, J. K. *et al.* Genome organization of membrane-containing bacteriophage PRD1. *Virology* **183**, 658–676 (1991).
 34. Rydman, P. S. & Bamford, D. H. Bacteriophage PRD1 DNA entry uses a viral membrane-associated transglycosylase activity. *Mol. Microbiol.* **37**, 356–363 (2000).
 35. Rydman, P. S. & Bamford, D. H. The lytic enzyme of bacteriophage PRD1 is associated with the viral membrane. *J. Bacteriol.* **184**, 104–110 (2002).
 36. Thuman-Commike, P. A. *et al.* Mechanism of scaffolding-directed virus assembly suggested by comparison of scaffolding-containing and scaffolding-lacking P22 procapsids. *Biophys. J.* **76**, 3267–3277 (1999).
 37. Morais, M. C. *et al.* Bacteriophage phi29 scaffolding protein gp7 before and after prohead assembly. *Nature Struct. Biol.* **10**, 572–576 (2003).
 38. Journet, L., Agrain, C., Broz, P. & Cornelis, G. R. The needle length of bacterial injectisomes is determined by a molecular ruler. *Science* **302**, 1757–1760 (2003).
 39. Burnett, R. M. The structure of the adenovirus capsid. II. The packing symmetry of hexon and its implications for viral architecture. *J. Mol. Biol.* **185**, 125–143 (1985).
 40. Sambrook, J. & Russell, D. W. *Molecular Cloning: a Laboratory Manual* 3rd edn (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2001).
 41. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
 42. Collaborative Computational Project, Number 4. The CCP 4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
 43. Brunger, A. T. *X-PLOR, Version 3.1: A system for X-ray crystallography and NMR* (Yale Univ. Press, New Haven, Connecticut, 1992).
 44. DeLano, W. L. & Brunger, A. T. The direct rotation function – rotational Patterson correlation search applied to molecular replacement. *Acta Crystallogr. D* **51**, 740–748 (1995).
 45. Diprose, J. M. *et al.* Translocation portals for the substrates and products of a viral transcription complex: the bluetongue virus core. *EMBO J.* **20**, 7229–7239 (2001).
 46. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
 47. Sonnhämmer, E. L. L., von Heijne, G. & Krogh, A. in *Proc. Sixth International Conference on Intelligent Systems for Molecular Biology* (ed. Glasgow, J.) 175–182 (AAAI Press, Menlo Park, California, 1998).
 48. Esnouf, R. M. Polyalanine reconstruction from C α positions using the program CALPHA can aid initial phasing of data by molecular replacement procedures. *Acta Crystallogr. D* **53**, 665–672 (1997).
 49. Frishman, D. & Argos, P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* **27**, 329–335 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M.-L. Perälä and S. Ollila for technical assistance, the staff at the European Synchrotron Radiation Facility for beamline support, and E. Mancini for assistance with data collection. This investigation was supported by research grants from the Academy of Finland to J.K.H.B. and S.J.B., research grants and the Finnish Centre of Excellence Program (2000–2005) from the Academy of Finland to D.H.B., a grant from the National Science Foundation to R.M.B., the Biotechnology and Biological Sciences Research Council and the Medical Research Council, UK and a grant from the Human Frontiers Science Program to D.I.S., R.M.B. and D.H.B. S.D.F. is supported by the Wellcome Trust, J.M.G. by the Royal Society and D.I.S. by the UK Medical Research Council.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.I.S. (dave@strubi.ox.ac.uk) or J.K.H.B. (jaana.bamford@helsinki.fi). Coordinates for proteins P3, P30, P31 and P16 within the icosahedral asymmetric unit have been deposited in the Protein Data Bank under accession code 1w8x.

High-energy particle acceleration in the shell of a supernova remnant

F. A. Aharonian¹, A. G. Akhperjanian², K.-M. Aye³, A. R. Bazer-Bachi⁴, M. Beilicke⁵, W. Benbow¹, D. Berge¹, P. Berghaus^{6*}, K. Bernlöhr^{1,7}, O. Bolz¹, C. Boisson⁸, C. Borgmeier⁷, F. Breitting⁷, A. M. Brown³, J. Bussons Gordo⁹, P. M. Chadwick³, V. R. Chitnis^{10,20*}, L.-M. Chounet¹¹, R. Cornils⁵, L. Costamante^{1,20}, B. Degrange¹¹, A. Djannati-Atai⁶, L. O'C. Drury¹², T. Ergin⁷, P. Espigat⁶, F. Feinstein⁹, P. Fleury¹¹, G. Fontaine¹¹, S. Funk¹, Y. A. Gallant⁹, B. Giebels¹¹, S. Gillesen¹, P. Goret¹³, J. Guy¹⁰, C. Hadjichristidis³, M. Hauser¹⁴, G. Heinzelmann⁵, D. Henri¹⁵, G. Hermann¹, J. A. Hinton¹, W. Hofmann¹, M. Holleran¹⁶, D. Horns¹, O. C. de Jager¹⁶, I. Jung^{1,14*}, B. Khélifi¹, Nu. Komin⁷, A. Konopelko^{1,7}, I. J. Latham³, R. Le Gallou³, M. Lemoine¹¹, A. Lemièrre⁶, N. Leroy¹¹, T. Lohse⁷, A. Marcowith⁴, C. Masterson^{1,20}, T. J. L. McComb³, M. de Naurois¹⁰, S. J. Nolan³, A. Noutsos³, K. J. Orford³, J. L. Osborne³, M. Ouchrif^{10,20}, M. Panter¹, G. Pelletier¹⁵, S. Pita⁶, M. Pohl^{17*}, G. Pühlhofer^{1,14}, M. Punch⁶, B. C. Raubenheimer¹⁶, M. Raue⁵, J. Raux¹⁰, S. M. Rayner³, I. Redondo^{11,20*}, A. Reimer¹⁷, O. Reimer¹⁷, J. Ripken⁵, M. Rivoal¹⁰, L. Rob¹⁸, L. Rolland¹⁰, G. Rowell¹, V. Sahakian², L. Saugé¹⁵, S. Schlenker⁷, R. Schlickeiser¹⁷, C. Schuster¹⁷, U. Schwanke⁷, M. Siewert¹⁷, H. Sol⁸, R. Steenkamp¹⁹, C. Stegmann⁷, J.-P. Tavernet¹⁰, C. G. Théoret⁶, M. Tluczykont^{11,20}, D. J. van der Walt¹⁶, G. Vasileiadis⁹, P. Vincent¹⁰, B. Visser¹⁶, H. J. Völk¹ & S. J. Wagner¹⁴

¹Max-Planck-Institut für Kernphysik, PO Box 103980, D 69029 Heidelberg, Germany

²Yerevan Physics Institute, 2 Alikhanian Brothers Street, 375036 Yerevan, Armenia

³Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK

⁴Centre d'Etude Spatiale des Rayonnements, CNRS/UPS, 9 av. du Colonel Roche, BP 4346, F-31029 Toulouse Cedex 4, France

⁵Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, D 22761 Hamburg, Germany

⁶Physique Corpusculaire et Cosmologie, IN2P3/CNRS, Collège de France, 11 Place Marcelin Berthelot, F-75231 Paris Cedex 05, France

⁷Institut für Physik, Humboldt-Universität zu Berlin, Newtonstrasse 15, D 12489 Berlin, Germany

⁸LUTH, UMR 8102 du CNRS, Observatoire de Paris, Section de Meudon, F-92195 Meudon Cedex, France

⁹Groupe d'Astroparticules de Montpellier, IN2P3/CNRS, Université Montpellier II, CC85, Place Eugène Bataillon, F-34095 Montpellier Cedex 5, France

¹⁰Laboratoire de Physique Nucléaire et de Hautes Energies, IN2P3/CNRS, Universités Paris VI & VII, 4 Place Jussieu, F-75231 Paris Cedex 05, France

¹¹Laboratoire Leprince-Ringuet, IN2P3/CNRS, Ecole Polytechnique, F-91128 Palaiseau, France

¹²Dublin Institute for Advanced Studies, 5 Merrion Square, Dublin 2, Ireland

¹³Service d'Astrophysique, DAPNIA/DSM/CEA, CE Saclay, F-91191 Gif-sur-Yvette, France

¹⁴Landessternwarte, Königstuhl, D 69117 Heidelberg, Germany

¹⁵Laboratoire d'Astrophysique de Grenoble, INSU/CNRS, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France

¹⁶Unit for Space Physics, North-West University, Potchefstroom 2520, South Africa

¹⁷Institut für Theoretische Physik, Lehrstuhl IV: Weltraum und Astrophysik, Ruhr-Universität Bochum, D 44780 Bochum, Germany

¹⁸Institute of Particle and Nuclear Physics, Charles University, V Holesovickach 2, 180 00 Prague 8, Czech Republic

¹⁹University of Namibia, Private Bag 13301, Windhoek, Namibia

²⁰European Associated Laboratory for Gamma-Ray Astronomy

* Present addresses: Université Libre de Bruxelles, Faculté des Sciences, Campus de la Plaine, CP230, Boulevard du Triomphe, 1050 Bruxelles, Belgium (P.B.); Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005, India (V.R.C.); Washington University, Department of Physics, 1 Brookings Drive, CB 1105, St Louis, Missouri 63130, USA (I.J.); Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011-3160, USA (M.P.); Department of Physics and Astronomy, University of Sheffield, The Hicks Building, Hounsfield Road, Sheffield S3 7RH, UK (I.R.)

A significant fraction of the energy density of the interstellar medium is in the form of high-energy charged particles (cosmic rays)¹. The origin of these particles remains uncertain. Although

it is generally accepted that the only sources capable of supplying the energy required to accelerate the bulk of Galactic cosmic rays are supernova explosions, and even though the mechanism of particle acceleration in expanding supernova remnant (SNR) shocks is thought to be well understood theoretically^{2,3}, unequivocal evidence for the production of high-energy particles in supernova shells has proven remarkably hard to find. Here we report on observations of the SNR RX J1713.7–3946 (G347.3–0.5), which was discovered by ROSAT⁴ in the X-ray spectrum and later claimed as a source of high-energy γ -rays^{5,6} of TeV energies (1 TeV = 10^{12} eV). We present a TeV γ -ray image of the SNR: the spatially resolved remnant has a shell morphology similar to that seen in X-rays, which demonstrates that very-high-energy particles are accelerated there. The energy spectrum indicates efficient acceleration of charged particles to energies beyond 100 TeV, consistent with current ideas of particle acceleration in young SNR shocks.

RX J1713.7–3946, together with several other southern hemisphere SNRs, is a prime target for observations with the High Energy Stereoscopic System (HESS), a new system of four imaging atmospheric Cherenkov telescopes located in the Khomas Highland of Namibia. HESS^{7,8} (we note that V. F. Hess discovered cosmic rays) exploits the most effective detection technique for very-high-energy γ -rays, namely, the imaging of Cherenkov light from air showers. This technique, which was pioneered by the Whipple collaboration⁹, makes use of the fact that whenever a high-energy γ -ray hits the Earth's atmosphere it is absorbed and initiates a cascade of interactions with air atoms, leading to the formation of a shower of secondary charged particles. Those travelling faster than the local speed of light in air emit Cherenkov radiation, which results in a brief flash of blue Cherenkov light detectable at ground level. By using a telescope with sufficient mirror area to collect enough of the faint light signal, and a fast camera with fine pixelation, one can image the shower and reconstruct from this image the direction and energy of the primary γ -ray. Combined with the approach of stereoscopic imaging of the cascade using a system of telescopes, as pioneered by the HEGRA collaboration¹⁰, this yields a very powerful technique for imaging and obtaining energy spectra of astronomical sources at TeV energies.

The HESS experiment is such a stereoscopic system, consisting of four 13-m-diameter telescopes¹¹ spaced at the corners of a square of side 120 m, each equipped with a 960-phototube camera¹² covering a large field of view of diameter 5° . Construction of the telescope system started in 2001; the full array was completed in December 2003 with the commissioning of the fourth telescope. HESS has an angular resolution of a few arc minutes, an effective energy range from 100 GeV to 10 TeV with energy resolution of 15–20% and a flux sensitivity approaching 10^{-13} erg cm⁻² s⁻¹. These characteristics, together with its southern hemisphere location, make HESS ideally suited for spectroscopic and morphological studies of Galactic plane sources such as RX J1713.7–3946, which is now the first SNR shell to be confirmed as a TeV source. TeV emission has also been reported from the remnant of SN 1006 (ref. 13), a result which our observations have not yet allowed us to confirm with HESS¹⁴, and from Cassiopeia A^{15,16} (a classical core-collapse SNR and the weakest TeV source yet reported) whose northern location makes it inaccessible to HESS.

Here we present results from observations of RX J1713.7–3946 performed between May and August 2003 during two phases of the construction and commissioning of HESS. In the first phase, two telescopes were operated independently, with stereoscopic event selection done offline using GPS time stamps to identify coincident events. During the second phase, also using two telescopes, coincident events were selected in hardware using an array level trigger. The total on-source observation time was 26 h; after run selection and dead time correction a data set corresponding to 18.1 live hours was used in this analysis. At the trigger level (for observation altitude

angles of 60–75°), the energy thresholds for the two configurations were 250 GeV (without the array level trigger) and 150 GeV (with the array level trigger). In this analysis, hard cuts were used to select only well reconstructed showers. This primarily served to drastically reduce the number of background cosmic-ray events, but it also homogenized these data taken with the two different configurations and improved the angular resolution. Thereby systematic errors are greatly reduced at the expense of a higher energy threshold (~ 800 GeV for the combined data set).

Figure 1 shows the resulting count map centred on RX J1713.7–3946. The SNR stands out from the residual charged cosmic-ray background with a significance of 20 standard deviations. The overall shell structure is clearly visible and coincides closely with that seen in X-rays (see Fig. 2). To our knowledge, this is the first case where the identification of an active (that is accelerating) celestial γ -ray source (as opposed to a passive cloud or density enhancement, merely penetrated by energetic particles which are accelerated elsewhere) can be based, not just on a positional coincidence, but on the image morphology. The overall flux above 1 TeV is $(1.46 \pm 0.17 \text{ (statistical)} \pm 0.37 \text{ (systematic)}) \times 10^{-7}$ photons $\text{m}^{-2} \text{s}^{-1}$, which corresponds to 66%

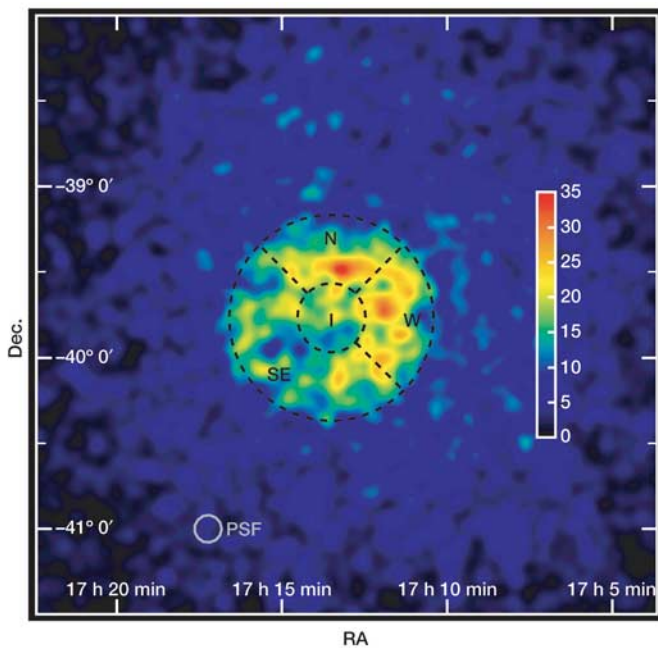


Figure 1 Wide field of view ($3.5^\circ \times 3.5^\circ$) around RX J1713.7 – 3946. Integration regions for flux estimates are shown. The flux above 1 TeV from the northern (N) rim is $(3.0 \pm 0.6) \times 10^{-8}$ photons $\text{m}^{-2} \text{s}^{-1}$, from the western (W) rim $(4.1 \pm 0.8) \times 10^{-8}$ photons $\text{m}^{-2} \text{s}^{-1}$, from the southeastern (SE) rim $(5.9 \pm 1.0) \times 10^{-8}$ photons $\text{m}^{-2} \text{s}^{-1}$ and the flux from the interior (I) $(1.7 \pm 0.6) \times 10^{-8}$ photons $\text{m}^{-2} \text{s}^{-1}$. The mean γ -ray brightnesses per unit solid angle (that is, the flux values normalized to the area) of the regions are in the ratio 1:1.4:1:1.2 (N:W:SE:I). These values might contradict the visual impression that the northwestern shell of the SNR is brighter. However, statistics of the data sample are limited, and the different areas were chosen for geometric reasons. Looking at the image, we see for example that dim regions are included in the seemingly brighter northern and western area, whereas the interior might gain from leakages from the northwestern shell. More detailed spatially resolved flux studies will have to await the advent of new data taken with the full HESS array with increased sensitivity. The 70% containment radius of the γ -ray point-spread function (PSF) for this data set with an energy threshold of 800 GeV is illustrated in the bottom left-hand corner (structures that are smaller than this circle should not be considered as real). The image is smoothed with a gaussian of standard deviation 3 arcmin (matched to the angular resolution of the instrument for this particular data set). Note that the efficiency of the camera falls off towards the edge of the field of view.

of the Crab nebula flux as measured by HESS. More elaborate analyses of these data using different background models (needed to determine the spectrum) and independent (and different) analysis chains confirm the results presented here.

The energy spectrum of the whole remnant is shown in Fig. 3. These data are well described by a power law (see Fig. 3 legend) with a photon index $\Gamma = 2.19 \pm 0.09 \pm 0.15$, as compared to the photon index of $2.84 \pm 0.15 \pm 0.20$ reported by the CANGAROO-II collaboration for the northwest part of the SNR⁶. The integral energy flux between 1 TeV and 10 TeV is estimated to be 3.5×10^{-11} erg $\text{cm}^{-2} \text{s}^{-1}$, which is an order of magnitude smaller than the non-thermal X-ray flux. More data will be taken with the full HESS array in 2004. The increased sensitivity (four instead of two telescopes) will enable spatially resolved spectral studies.

RX J1713.7–3946 (situated in the Galactic plane, in the constellation Scorpius) is one of the brighter Galactic X-ray SNRs¹⁷, with a flux density of a few times 10^{-10} erg $\text{cm}^{-2} \text{s}^{-1}$. In X-rays, it exhibits typical shell morphology, but remarkably the X-ray spectrum is completely dominated by a non-thermal continuum with no detectable line emission. The most plausible origin of these X-rays is the synchrotron radiation of 100-TeV electrons^{18,19}. However, because alternative explanations are not absolutely ruled out²⁰, only the detection of TeV emission from this SNR will provide unambiguous evidence for acceleration of particles to multi-TeV energies. Another important point is that the TeV signal may contain a component due to accelerated protons interacting with the ambient gas, as has been predicted theoretically^{2,3}. The contribution of this component should be significantly enhanced when the supernova shell overtakes nearby dense molecular clouds²¹, as seems to be the case for this object. The CO data²² suggest that a cloud is interacting with the northwestern part of the SNR, where a

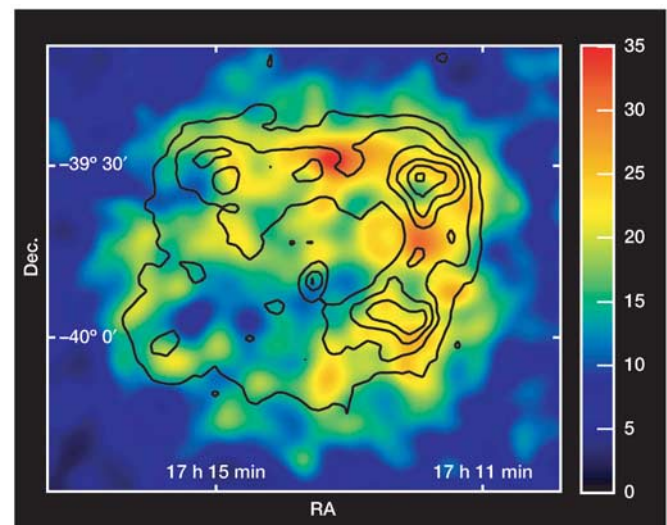


Figure 2 γ -ray image of the SNR RX J1713.7 – 3946 obtained with the HESS telescopes. Hard cuts were applied to select well-reconstructed γ -like events above 800 GeV. The map is smoothed as in Fig. 1, having the same scale in units of counts. The linear colour scale is in units of counts. We note that no background subtraction or camera-efficiency corrections have been applied. This demonstrates that the structures seen are not artefacts of the analysis but real and visible in the raw post-cuts data (the background in the field of view is at a level of about five counts, and the efficiency across the SNR changes by less than 10%). This image, obtained with a partial array during construction, demonstrates the ability of HESS to map extended objects. The superimposed (linearly spaced) contours show the X-ray surface brightness as seen by ASCA in the 1–3 keV range for comparison²⁵. Note that the angular resolution of ASCA is comparable to that of HESS which enables direct comparison of the two images. RA, right ascension; dec., declination.

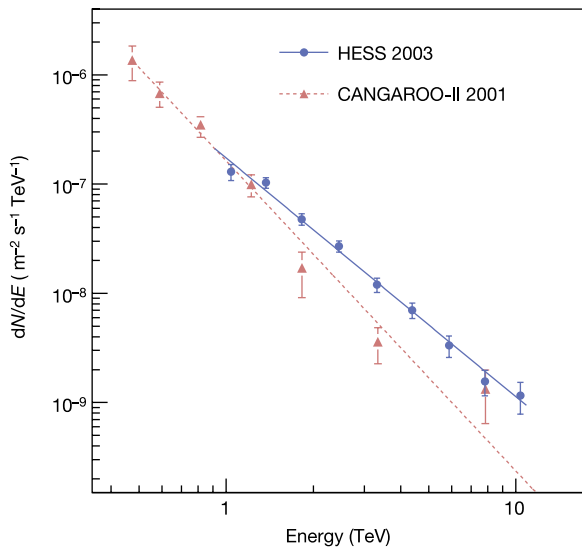


Figure 3 γ -ray energy spectrum of RX J1713.7 – 3946 as measured with the HESS telescopes. These data (blue filled circles) can be described by a power law, $dN/dE \propto E^{-\Gamma}$; the best fit result (blue solid line) gives $\Gamma = 2.19 \pm 0.09$ (statistical) ± 0.15 (systematic) with $\chi^2 = 5.9$ with 7 degrees of freedom. The integral flux above 1 TeV is found to be $(1.46 \pm 0.17 \text{ (statistical)} \pm 0.37 \text{ (systematic)}) \times 10^{-7} \text{ photons m}^{-2} \text{ s}^{-1}$. There is clearly no evidence for a cut-off in the data, but if one nevertheless attempts to fit an exponentially cut-off power law of the form $dN/dE \propto E^{-\Gamma} e^{-E/E_c}$, the minimum acceptable value of E_c is 4 TeV with a very hard photon index of $\Gamma_c = 1.5$. The spectral data points reported by the CANGAROO-II collaboration⁶ for the northwestern part of the remnant are shown for comparison as red triangles, and the best fit result as a dashed red line. The fact that CANGAROO-II reported a spectrum only for a part of the SNR prohibits at this stage a definite statement about the compatibility of the two measurements. Error bars, 1σ statistical errors.

striking spatial coincidence between the CO density peaks and the regions of peak X-ray emission is seen. The X-ray data²³ also indicate significant absorption column densities in the western part of the remnant, at values about twice those to the east. These indications fit qualitatively with the γ -ray image presented here, where TeV emission is seen from the whole SNR shell but with an increased flux from the northwestern side.

Given the recent estimates^{22–24} of the distance to the source of 1 kpc, if a significant part of the TeV flux were to be formed by interactions of cosmic-ray nuclei with gas atoms in the cloud with density n exceeding 100 cm^{-3} , the energetics implied by the γ -ray flux and the spectrum would be a few times $10^{49} n^{-1} \text{ erg}$ between 10 and 100 TeV. This is consistent with the picture of an SNR origin of Galactic cosmic rays involving about 10% efficiency for conversion of the mechanical energy of the explosion into non-thermal particles, and a production spectrum in the SNR which is approximately an E^{-2} power law from several GeV to about one PeV. Moreover, the γ -ray morphology is qualitatively what one would expect from particles accelerated at the shock interacting and radiating in the compressed post-shock region. The extension of the γ -ray spectrum up to energies of 10 TeV (see Fig. 3) requires an extremely effective accelerator boosting particles up to energies of at least 100 TeV. RX J1713.7 – 3946 is a complex object interacting with molecular clouds of different densities where the TeV emission might emerge from various processes. Without doubt there will be a contribution from energetic electrons through the inverse Compton process, especially from low-density regions (as in the eastern part) of the SNR. At the elevated densities likely to exist in the northwestern rim, π^0 -decays following proton–proton interactions, but also non-thermal Bremsstrahlung of electrons, could make significant contributions. Although disentangling the relative

contributions of the various processes is difficult, it should be possible through spatially resolved multi-wavelength studies—which will be undertaken using the full HESS array.

The multi-TeV image of a shell-type SNR presented here constitutes a significant step forward towards a solution of the long-standing puzzle of the origin of Galactic cosmic rays, as well as demonstrating an astronomical imaging technique operating at photon energies some 12 decades higher than those of visible light. $\square$

Received 23 June; accepted 17 August 2004; doi:10.1038/nature02960.

- Longair, M. S. *High Energy Astrophysics Vol. 2, Stars, the Galaxy and the Interstellar Medium* 2nd edn (Cambridge Univ. Press, Cambridge, 2002).
- Drury, L. O'C., Aharonian, F. A. & Völk, H. J. The gamma-ray visibility of supernova remnants. A test of cosmic-ray origin. *Astron. Astrophys.* **287**, 959–971 (1994).
- Naito, T. & Takahara, F. High-energy gamma-ray emission from supernova remnants. *J. Phys. G* **20**, 477–486 (1994).
- Pfeffermann, E. & Aschenbach, B. in *Röntgenstrahlung from the Universe* (eds Zimmermann, H. U., Truemper, J. E. & Yorke, H.) 267–268 (Report 263, MPE, Garching, 1996).
- Muraishi, H. *et al.* Evidence for TeV gamma-ray emission from the shell type SNR RX J1713.7–3946. *Astron. Astrophys.* **354**, L57–L61 (2000).
- Enomoto, R. *et al.* The acceleration of cosmic-ray protons in the supernova remnant RX J1713.7–3946. *Nature* **416**, 823–826 (2002).
- Hinton, J. A. (H.E.S.S. Collaboration) The status of the HESS project. *N. Astron. Rev.* **48**, 331–337 (2004).
- Hofmann, W. (H.E.S.S. Collaboration) in *Proc. 28th ICRC (Tsukuba)* (eds Kajita, T., Asaoka, Y., Kawachi, A., Matsubara, Y. & Sasaki, M.) 2811–2814 (Univ. Academy Press, Tokyo, 2003).
- Weekes, T. C. *et al.* Observation of TeV gamma rays from the Crab nebula using the atmospheric Cherenkov imaging technique. *Astrophys. J.* **342**, 379–395 (1989).
- Aharonian, F. A. *et al.* The time averaged TeV energy spectrum of Mkn 501 of the extraordinary 1997 outburst as measured with the stereoscopic Cherenkov telescope system of HEGRA. *Astron. Astrophys.* **349**, 11–28 (1999).
- Bernlöhr, K. *et al.* The optical system of the H.E.S.S. imaging atmospheric Cherenkov telescopes, Part I: layout and components of the system. *Astropart. Phys.* **20**, 111–128 (2003).
- Vincent, P. *et al.* *Proc. 28th ICRC (Tsukuba)* (eds Kajita, T., Asaoka, Y., Kawachi, A., Matsubara, Y. & Sasaki, M.) 2887–2890 (Univ. Academy Press, Tokyo, 2003).
- Tanimori, T. *et al.* Discovery of TeV gamma rays from SN 1006: Further evidence for the supernova remnant origin of cosmic rays. *Astrophys. J.* **497**, L25–L28 (1998).
- Masterson, C. (H.E.S.S. Collaboration) in *Proc. 28th ICRC (Tsukuba)* (eds Kajita, T., Asaoka, Y., Kawachi, A., Matsubara, Y. & Sasaki, M.) 2323–2326 (Univ. Academy Press, Tokyo, 2003).
- Aharonian, F. A. *et al.* Evidence for TeV gamma ray emission from Cassiopeia A. *Astron. Astrophys.* **370**, 112–120 (2001).
- Berezhko, E. G., Pühlhofer, G. & Völk, H. J. Gamma-ray emission from Cassiopeia A produced by accelerated cosmic rays. *Astron. Astrophys.* **400**, 971–980 (2003).
- Slane, P. *et al.* Nonthermal X-ray emission from the shell-type supernova remnant G347.3–0.5. *Astrophys. J.* **525**, 357–367 (1999).
- Ellison, D. C., Slane, P. & Gaensler, P. M. Broadband observations and modeling of the shell-type supernova remnant G347.3–0.5. *Astrophys. J.* **563**, 191–201 (2001).
- Uchiyama, Y., Aharonian, F. A. & Takahashi, T. Fine-structure in the nonthermal X-ray emission of SNR RX J1713.7–3946 revealed by Chandra. *Astron. Astrophys.* **400**, 567–574 (2003).
- Laming, J. M. A thermal model for the featureless X-ray emission from SN 1006? *Astrophys. J.* **499**, 309–314 (1998).
- Aharonian, F. A., Drury, L. O'C. & Völk, H. J. GeV/TeV gamma-ray emission from dense molecular clouds overtaken by supernova shells. *Astron. Astrophys.* **285**, 645–647 (1994).
- Fukui, Y. *et al.* Discovery of interacting molecular gas toward the TeV gamma-ray peak of the SNR G 347.3–0.5. *Publ. Astron. Soc. Jpn* **55**, L61–L64 (2003).
- Cassam-Chenaï, A. *et al.* XMM-Newton observations of the supernova remnant RX J1713.7–3946 and its central source. *Astron. Astrophys.* (in the press).
- Koo, B., Kang, J. & McClure-Griffiths, N. in *Young Neutron Stars and Their Environments* (eds Camilo, F. & Gaensler, B. M.) *IAU Symp.* 85–86 (Astronomical Society of the Pacific, San Francisco, 2004).
- Uchiyama, Y., Takahashi, T. & Aharonian, F. A. Flat spectrum X-ray emission from the direction of a molecular cloud associated with SNR RX J1713.7–3946. *Publ. Astron. Soc. Jpn* **54**, L73–L77 (2002).

Acknowledgements The support of the Namibian authorities and of the University of Namibia in facilitating the construction and operation of HESS is acknowledged. We also thank the following for support: the German Ministry for Education and Research (BMBF), the Max Planck Society, the French Ministry for Research, the CNRS-IN2P3 and the Astroparticle Interdisciplinary Programme of the CNRS, the UK Particle Physics and Astronomy Research Council (PPARC), the IPNP of Charles University, the South African Department of Science and Technology and National Research Foundation, and the University of Namibia. The European Associated Laboratory for Gamma-Ray Astronomy is jointly supported by CNRS and MPG. We appreciate the work of the technical support staff in Berlin, Durham, Hamburg, Heidelberg, Palaiseau, Paris, Saclay and Namibia in the construction and operation of the equipment. We also thank Y. Uchiyama for supplying the ASCA X-ray data shown in Fig. 2.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.B. (David.Berge@mpi-hd.mpg.de).

An interplanetary shock traced by planetary auroral storms from the Sun to Saturn

Renée Prangé¹, Laurent Pallier¹, Kenneth C. Hansen², Russ Howard³, Angelos Vourlidas³, Régis Courtin¹ & Chris Parkinson⁴

¹LESIA, Observatoire de Paris, 5 place Jules Janssen, 92195 Meudon, France

²Department of Atmospheric, Oceanic and Space Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA

³Naval Research Laboratory, 4555 Overlook Avenue, SW Washington, DC 20375, USA

⁴Californian Institute of Technology, Jet Propulsion Laboratory and the NASA Astrobiology Institute, MS150-21 E. California Boulevard, Pasadena, California 91125, USA

A relationship between solar activity and aurorae on Earth was postulated^{1,2} long before space probes directly detected plasma propagating outwards from the Sun³. Violent solar eruption events trigger interplanetary shocks⁴ that compress Earth's magnetosphere, leading to increased energetic particle precipitation into the ionosphere and subsequent auroral storms^{5,6}. Monitoring shocks is now part of the 'Space Weather' forecast programme aimed at predicting solar-activity-related environmental hazards. The outer planets also experience aurorae, and here we report the discovery of a strong transient polar emission on Saturn, tentatively attributed to the passage of an interplanetary shock—and ultimately to a series of solar coronal mass ejection (CME) events. We could trace the shock-triggered events from Earth, where auroral storms were recorded, to Jupiter, where the auroral activity was strongly enhanced, and to Saturn, where it activated the unusual polar source. This establishes that shocks retain their properties and their ability to trigger planetary auroral activity throughout the Solar System. Our results also reveal differences in the planetary auroral responses on the passing shock, especially in their latitudinal and local time dependences.

As for that of the Earth, outer-planet magnetospheres—the domain where the magnetic pressure of planetary origin exceeds the solar-wind kinetic (ram) pressure—are shaped by the solar-wind flow and they also display permanent aurorae near the magnetic poles at infrared, far-ultraviolet (FUV) and radio wavelengths^{7–11}. Similarly, when a high-ram-pressure interplanetary shock wave reaches the planet, the magnetosphere is suddenly compressed. Among other consequences, the strength and distribution of currents and plasma are altered, and enhanced fluxes of energetic charged particles are expected to precipitate into the auroral ovals, which must be strongly activated. Indeed, strong statistical correlations were found between auroral radio emissions and the solar-wind density and ram pressures^{12–16}, particularly for the Earth and Saturn, whose magnetospheres are dominated by interactions with the solar wind at the polar-cap boundary¹⁷, between magnetic field lines open to the interplanetary magnetic field (IMF) and closed field lines. The correlation is looser at Jupiter, owing to the complex nature of its magnetosphere, partly controlled by processes of planetary origin^{18–20}. For the Earth, an extended set of observations of the Sun, coordinated with measurements of the solar-wind physical parameters (density, velocity, composition, magnetic field strength and direction, and so on) and with various indicators of the auroral activity (among which are radio, visible and ultraviolet emissions), allows predictions of the ionosphere and atmosphere response to any interplanetary shock and or change in the direction of the IMF. This is not the case for the outer planets where there are no such coordinated observations, except during the Cassini fly-by

of Jupiter in late 2000, where three occurrences of auroral radio storms coincided with the arrival of interplanetary shocks at the planet²¹. In the absence of any image, the nature and location of the auroral events could not be compared with that of geomagnetic storms.

On 7 and 8 December 2000, FUV high-spatial-resolution images of Saturn were taken with the Hubble space telescope, HST (Fig. 1) with an excellent view of the southern polar region. The auroral oval morphology was very different from one day to the next. On 8 December, a narrow oval surrounding the pole at high latitude, and assigned to large-scale currents at the boundary between open and closed field lines¹⁷, resembles the few Saturn aurorae observed earlier²²; this configuration may be considered as a steady-state reference. By contrast, the aurora observed on the previous day is of

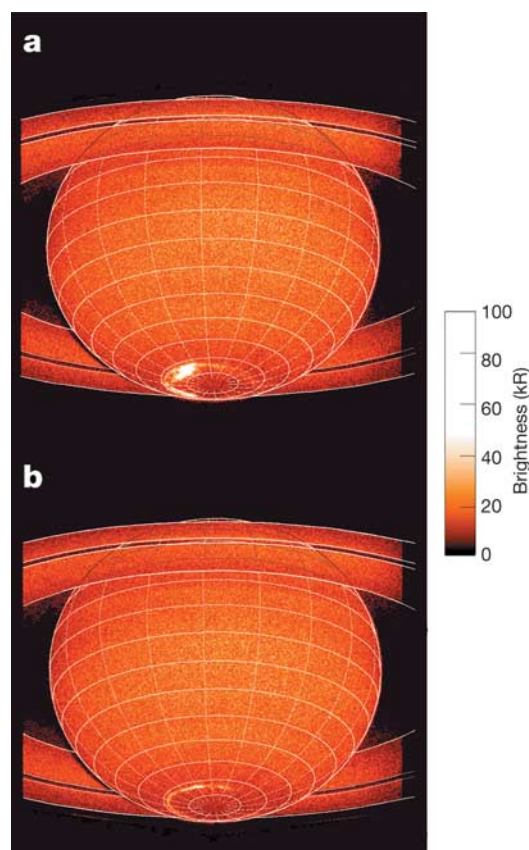


Figure 1 Discovery of an auroral storm on Saturn. The images, in a false-orange brightness colour code, were taken with the Space Telescope Imaging Spectrograph (STIS) onboard HST at 11:30 UT on 7 December 2000 (**a**), and 10:00 UT on 8 December (**b**). The exposure time is 480 s, during which Saturn rotated by $\sim 5^\circ$. A spectral range including H_2 Werner–Lyman auroral bands between ~ 130 and ~ 165 nm is isolated by selecting the filter SRF2 in combination with the FUV MAMA solar blind detector. Nevertheless, some solar light is reflected by the planet disc in the FUV (up to ~ 190 nm). It is used to define the exact location of the planet on the detector and hence to correct pointing inaccuracies. After fitting an oblate spheroid plus ring system to the image, a grid of coordinates, with steps of 10° in latitude and 20° in longitude, is overlaid to determine the coordinates of the features of interest. The accuracy of the limb-fitting procedure is $\sim \pm 2$ pixels in each direction. The pixel size is 0.024×0.024 arcsec² and the field of view is 24.7×24.7 arcsec². The rotational phase of Saturn at the time of the observations is characterized by the longitude of the sub-Earth point, or local noon (SEP = 243° and 283° respectively in the images used here). The north polar region is hidden behind the ring system owing to the large tilt of Saturn's rotation axis at that time. Conversely, the whole region limited by the southern auroral oval is visible from Earth. Note the similarity in geometry and the differences in nightside brightness between the auroral oval on both days, but especially the very bright feature visible poleward of the dawnside oval on 7 December. This feature has totally disappeared on 8 December.

a kind never observed before: a similar oval is also present, but a very bright feature has transiently developed inside, on polar-cap open field lines. This suggests that we have witnessed for the first time an auroral storm at Saturn, which may be related to an as-yet-unknown interaction of the solar wind with Saturn's magnetosphere.

At that time, the Sun, the Earth, Jupiter and Saturn were nearly aligned (Fig. 2). Hence, any interplanetary shock would successively encounter the three planets along its radial propagation outward, even if its angular extent were small. A three-month observing campaign of Jupiter was also in progress, involving the Galileo spacecraft in the jovian magnetosphere, Cassini in the nearby solar wind, and the HST remotely imaging. On the other hand, observations of the Sun and of Earth's aurora are performed on a continuous basis (for example, by the POLAR and IMAGE Earth orbiters), together with solar-wind measurements in the near vicinity of the Earth (for example, by the ACE and WIND spacecraft), to predict environmental hazards due to energetic particle precipitation into the auroral ionosphere (the 'Space Weather' programmes). A crude kinetic retracing of the solar-wind plasma suggests that if any shock had triggered the auroral storm on Saturn, it should have formed in early November near the Sun. Indeed, the LASCO coronagraphic spectrograph onboard the Solar and Heliospheric Observatory (SOHO) detected a series of CMEs, five of them being directed toward the Earth over the 1–10 November period.

Thus all the conditions except one have been met that are required to investigate the interaction of an interplanetary shock with the Earth, Jupiter and Saturn consecutively. Only solar-wind measurements near Saturn were still needed to characterize the passage of a shock directly.

We could possibly have extrapolated them from the Cassini measurements near Jupiter, that is, 4 AU upstream (1 astronomical unit, AU = 149.5×10^6 km) using a mere ballistic projection, $t_1 = t_0 + \Delta r/v_{sw}$ and correcting for the difference in heliocentric longitude. However, this procedure, based on a purely 'kinematic' approach and ignoring plasma physics considerations, becomes more and more inaccurate as distance increases.

Therefore, we developed a more self-consistent method for obtaining the solar-wind conditions at the outer planets. We modelled the radial evolution of the plasma with a one-dimensional

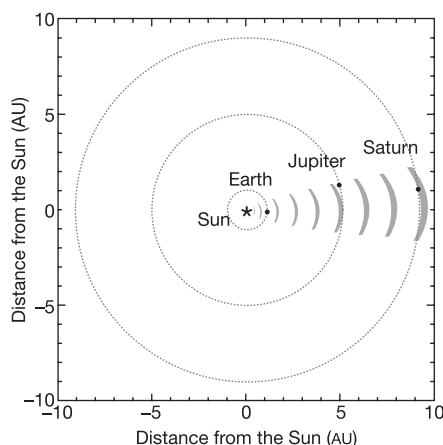


Figure 2 Geometrical configuration of the planets in late 2000. The relative heliocentric longitudes of the Earth, Jupiter and Saturn at the time when the interplanetary shock discussed in this study (in grey, schematic) passed each of them, are plotted. At the beginning (the end) of the passage, the angle from the Sun–Jupiter direction to the Sun–Earth direction is 23° (17.5°) as counted clockwise. For Saturn, these angles are 15° and 9° , respectively. Consequently, the code used to propagate the solar wind is anticipated to be reasonably accurate at Jupiter, and more accurate at Saturn (see text).

spherical magnetohydrodynamic code, using measurements made at Earth orbit as input. It was derived from the Versatile Advection Code²³, which was modified to accept the ACE data as input at 1 AU, the inner boundary of the simulations. Although the assumption of spherical symmetry is not strictly valid, the comparison of the propagated solar wind with Cassini measurements shows a good match, particularly when the planet is close to alignment with the Sun–Earth direction²⁴.

The 1–10 November series of CMEs triggered a series of five shocks, detected near the Earth (1 AU) by WIND and ACE about two days later. During this period, the POLAR Earth orbiter recorded correlated phases of geomagnetic storms and relaxation phases.

The code predicts that the individual shocks observed at 1 AU later merged into a single long-duration disturbance (fast shocks can overtake slower ones). This disturbance is predicted to have passed the jovian magnetosphere in the 18–24 November timeframe, coinciding with the first of the jovian radio auroral storms mentioned above²¹. Compared with the Cassini *in-situ* data, the model shock arrives ~ 0.7 days too early. This small disagreement stems almost entirely from a combination of the IMF Parker spiral structure with the $\sim 20^\circ$ Earth–Jupiter difference in heliospheric longitude²⁴. For Saturn, with a longitude difference of only $\sim 10^\circ$, this effect is expected to be smaller, and hence the predictions more accurate.

We used the code, validated at Jupiter, to propagate the shock further out, and we estimate that it passed Saturn between early 2 December and the morning of 8 December. Consequently, our first HST image exhibiting the polar storm was taken near the end of the shock interaction with Saturn's magnetosphere, while the second more 'typical' image is representative of a relaxation phase toward normal conditions, under a quiet ambient solar wind.

Thus we have assembled the first such synoptic view of the propagation of a CME-driven interplanetary shock from the Sun to Saturn, from *in-situ* measurements at the Earth and Jupiter (1 and 5 AU), and model-propagated plasma parameters at Jupiter and Saturn (5 and 9 AU). At each planet, a strong auroral response is recorded, based on POLAR images at the Earth, on Cassini radio observations at Jupiter, and on HST FUV images at Saturn. Figure 3 summarizes the observations. The solar-wind magnetic field strength and plasma ram pressure have been plotted in consecutive panels at the three planets. In parallel we display the last disturbed auroral image and the first 'relaxed' one for the Earth and Saturn. For Jupiter, we have instead plotted the radio output response.

We now compare the detailed characteristics of the auroral storms at the three planets, to progress towards an understanding of the physical processes at work during interactions between shocks and planetary magnetospheres.

At Earth, the quiet aurora, under solar-wind control, is located on the night side on closed magnetotail field lines. Compression of the magnetosphere by a shock leads to dramatic shifts towards low latitudes and brightenings of the auroral ovals (by up to three orders of magnitudes in extreme cases), which expand in local time both sides of midnight²⁵ (see Fig. 3). Increased emission from the footprint of the dayside polar cusps, where 'direct' entry of solar-wind plasma takes place, can also be observed as a result of reconnection events with the IMF and plasma injections²⁶.

The Saturn 'normal' aurora also suggests some kind of solar-wind-related local-time organization of the steady-state aurora. But in contrast with the Earth, the asymmetry in latitude and brightness develops between dawn and dusk¹⁷. During the storm, the auroral oval remains almost identical in size and location, but it is globally brighter by 50% (from 1.5×10^{10} to 2.2×10^{10} watts of total flux radiated in the H and H₂ FUV spectrum). The brightness distribution along the oval also varies, its midnight-to-dawn sector (01:00–05:30 local time) becoming very active, up to 70 kR (1 R = 10^6 photons radiated in 4π steradians), almost four times its next-day brightness.

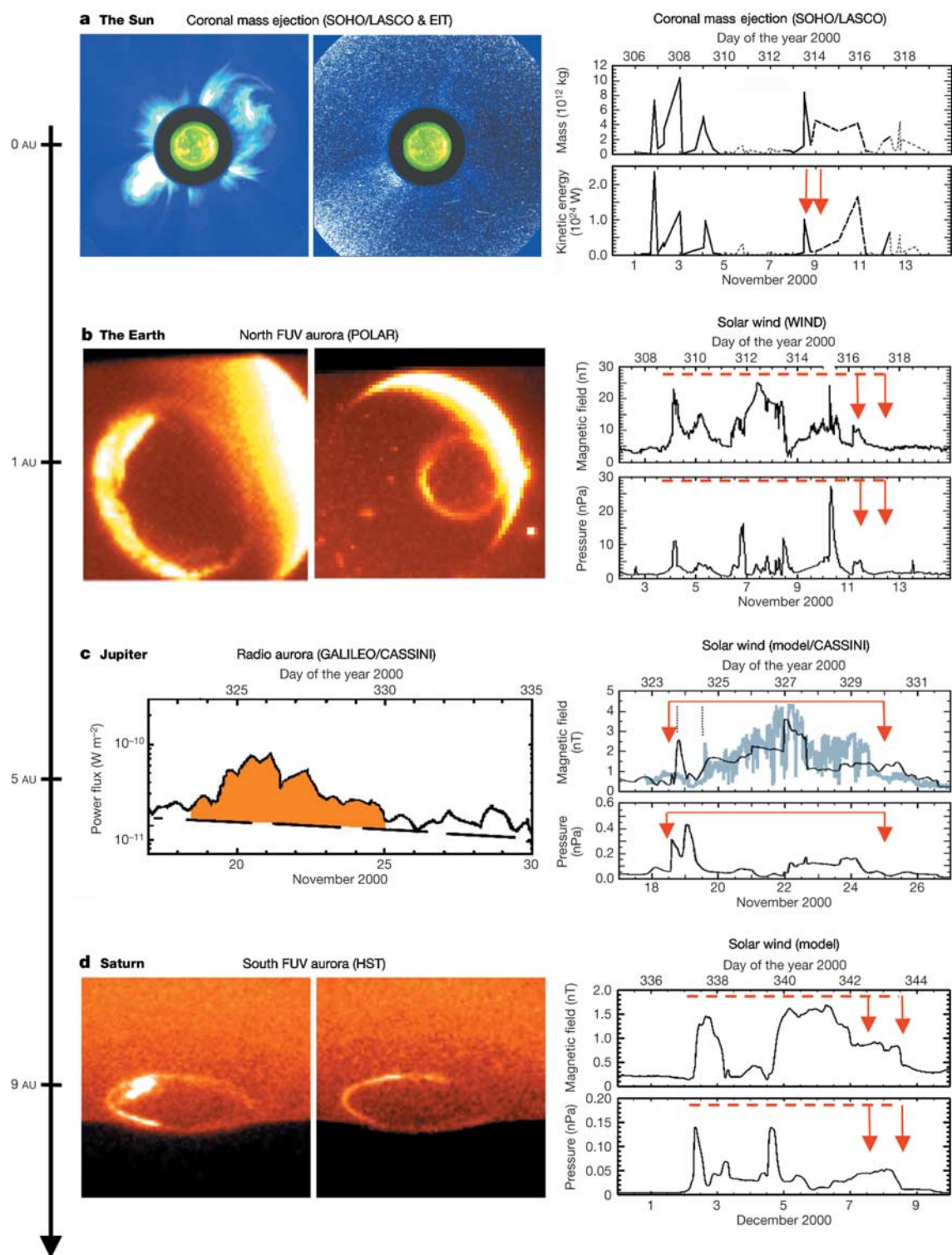


Figure 3 Synoptic view of the CME-driven shock propagation and auroral consequences. **a**, the Sun; **b**, the Earth; **c**, Jupiter; **d**, Saturn. Right column, solar-wind plasma parameters as a function of distance from the Sun (0 AU) to Saturn (9 AU). For the CME, the mass m and the kinetic energy $E_c = 1/2mv^2$ of the ejection event are plotted. For the interplanetary shocks, the magnetic field strength B , and the ram pressure $p = \rho v^2$. The dashed red lines and arrows indicate the duration of the event and the date of the images on the left. In **a**, thick solid lines, thick dashed lines, and thin dashed lines are for CMEs for which the geometry allows an encounter with the planets, marginally allows it, and does not allow it, respectively. In **c**, the magnetohydrodynamic-model propagated

shock (solid line) is compared with the local Cassini measurements (thick grey line): the dotted vertical lines around 18 November highlight the small difference (~ 0.7 days) between the shock onset from the model and from the data. Left column, SOHO images of the solar CME (green brightness scale for EIT images of the solar disc, and blue code for LASCO coronagraphic images of the neighbouring solar wind) and of the planetary auroral responses to the shock (orange brightness scale). The Earth aurora has been recorded by the visible/FUV imager onboard POLAR (<http://eiger.physics.uiowa.edu/~vis/images/>). Because simultaneous auroral images are not available at Jupiter, we have displayed instead the global response at radio wavelengths¹⁹.

In addition, there is a very bright, physically disconnected, feature inside the dawnside oval, at the footprint of polar-cap field lines. It extends polewards up to $\sim 78^\circ\text{S}$ ($\sim 6^\circ$ poleward of the oval), and is confined to the 5:30–11:00 local time sector. Its peak brightness (95 kR) largely exceeds that of the oval and its total auroral output, 2.4×10^{10} watts, is comparable to the oval output despite its very limited spatial extent.

This suggests that, in response to the passage of the shock, the auroral oval mainly brightens at both planets on the night side, at the footprint of magnetotail magnetic field lines. However, Saturn does not exhibit the expansion towards lower latitudes that is typical of the geomagnetic storm oval. As for the bright polar-cap feature, it does not seem to have any terrestrial counterpart. It is reminiscent of the terrestrial polar cusp, frequently observed during dayside reconnection of closed field lines with the IMF, on or slightly poleward of the oval, and which is detected as a variable feature on Jupiter as well²⁶. But, in contrast to the saturnian feature, the terrestrial polar cusp is close to noon²⁷.

Recent studies, however, indicate that the IMF By east–west component (that is, dawn–dusk on Saturn) has a strong influence on the local-time position of the cusp and may shift it towards morning or afternoon by as much as several hours^{28,29}. Even though polar cusps on Earth have not been observed as far as 6:00 local time, this is an issue that should be investigated. Unfortunately, the direction of the IMF cannot be extrapolated as the other solar-wind properties can be. Another interpretation could be that strong reconnection at Saturn does not occur on the day side, but rather on the morning flank of the magnetopause, where it has been suggested that a Kelvin–Helmholtz instability triggers the saturnian auroral radio emissions³⁰. It will be possible to check these hypotheses with local plasma measurements taken at Saturn by the Cassini mission. □

Received 14 May; accepted 27 August 2004; doi:10.1038/nature02986.

- de Mairan, J. J. *Traité Physique et Historique de l'Aurore Boréale* 245–279 (Mémoire de l'Académie Royale des Sciences, Imprimerie Royale, Paris, 1733).
- Chapman, S. & Ferraro, V. C. A. A new theory of magnetic storms. *Nature* **126**, 129 (1930); *Terr. Magn. Atmos. Electr.* **36**, 171–186 (1931); *Terr. Magn. Atmos. Electr.* **37**, 147–156 (1932).
- Licht, A. L. Properties of the solar wind during sunspot minimum. *J. Geophys. Res.* **65**, 1397–1400 (1960).
- Greenstadt, E. W. Interplanetary magnetic effects of solar flares: Explorer 18 and Pioneer 5. *J. Geophys. Res.* **70**, 5451–5452 (1965).
- Snyder, C. W., Neugebauer, M. & Rao, U. R. The solar wind velocity and its correlation with cosmic-ray variations and with solar and geomagnetic activity. *J. Geophys. Res.* **68**, 6361–6370 (1963).
- Axford, W. I. The interaction between the solar wind and the earth's magnetosphere. *J. Geophys. Res.* **67**, 3791–3796 (1962).
- Carr, T. D. Jupiter's decametric rotation period and the source A emission beam. *Phys. Earth Planet. Inter.* **6**, 21–28 (1972).
- Zarka, P. Auroral radio emissions at the outer planets: Observations and theories. *J. Geophys. Res.* **103**, 20159–20194 (1998).
- Broadfoot, A. L. *et al.* Extreme ultraviolet observations from Voyager 1 encounter with Jupiter. *Science* **204**, 979–982 (1979).
- Clarke, J. T. *et al.* Spatial imaging of hydrogen Lyman-alpha emission from Jupiter. *Astrophys. J.* **240**, 696–701 (1980).
- Broadfoot, A. L. *et al.* Extreme ultraviolet observations from Voyager 1 encounter with Saturn. *Science* **212**, 206–211 (1981).
- Gallagher, D. L. & D'Angelo, N. Correlation between solar wind parameters and auroral kilometric radiation intensity. *Geophys. Res. Lett.* **8**, 1087–1089 (1981).
- Zarka, P. & Genova, F. Low-frequency jovian emission and solar wind magnetic sector structure. *Nature* **306**, 767–768 (1983).
- Desch, M. D. & Barrow, C. H. Direct evidence for solar wind control of Jupiter's hectometer wavelength radio emission. *J. Geophys. Res.* **89**, 6819–6823 (1984).
- Rabl, G. K. F. Periodicities of jovian broad-band kilometric radiation observed by Ulysses. *Planet. Space Sci.* **41**, 1079–1083 (1993).
- Desch, M. D. & Rucker, H. O. The relationship between Saturn kilometric radiation and the solar wind. *J. Geophys. Res.* **88**, 8999–9006 (1983).
- Cowley, S. H. W., Bunce, E. J. & Prangé, R. Saturn's polar ionospheric flows and their relation to the main auroral oval. *Ann. Geophys.* **22**, 2685–2688 (2004).
- Vasyliunas, V. M. in *Physics of the Jovian Magnetosphere* (ed. Dessler, A. J.) 395–453 (Cambridge Univ. Press, Cambridge, UK, 1990).
- Prangé, R. *et al.* Detailed study of FUV Jovian auroral features with the post COSTAR Hubble Faint Object Camera. *J. Geophys. Res.* **103**, 20195–20210 (1998).
- Pallier, L. & Prangé, R. More about the structure of the high latitude Jovian aurorae. *Planet. Space Sci.* **49**, 1159–1173 (2001).
- Gurnett, D. A. *et al.* Control of Jupiter's radio emission and aurorae by the solar wind. *Nature* **415**, 985–987 (2002).

- Trauger, J. *et al.* Saturn's hydrogen aurora: Wide field and planetary camera 2 imaging from the Hubble Space Telescope. *J. Geophys. Res.* **103**, 20237–20244 (1998).
- Tóth, G. *Versatile Advection Code*; code available at (<http://www.phys.uu.nl/~toth/>) (1994).
- Hanlon, P. G. *et al.* On the evolution of the solar wind between 1 and 5 AU at the time of the Cassini Jupiter flyby: Multispacecraft observations of interplanetary coronal mass ejections including the formation of a merged interaction region. *J. Geophys. Res.* **109**, doi:10.1029/2003JA010112 (2004).
- Frank, L. A. & Craven, J. D. Imaging results from Dynamics Explorer. *Rev. Geophys.* **26**, 249–283 (1988).
- Pallier, L. & Prangé, R. Detection of the southern counterpart of the jovian northern polar cusp. Shared properties. *Geophys. Res. Lett.* **31**, doi:10.1029/2003GL018041 (2004).
- Lockwood, M. *et al.* The ionospheric signature of flux transfer events and solar wind dynamic pressure changes. *J. Geophys. Res.* **95**, 17113–17135 (1990).
- Sandholt, P. E., Farrugia, C. J., Moen, J., Cowley, S. W. H. & Lybäck, B. Dynamics of the aurora and associated currents during a cusp bifurcated event. *Geophys. Res. Lett.* **25**, 4313–4316 (1981).
- Milan, S. E., Lester, M., Cowley, S. W. H. & Brittnacher, M. Dayside convection and auroral morphology during an interval of northward interplanetary magnetic field. *Ann. Geophys.* **18**, 436–444 (2004).
- Galopeau, P. H. M., Zarka, P. & Le Quéau, D. Source location of Saturn's kilometric radiation: The Kelvin–Helmholtz instability hypothesis. *J. Geophys. Res.* **100**, 26337–26410 (1998).

Acknowledgements This work is based partly on observations with the NASA/ESA HST obtained at the STScI, which is operated by the AURA, Inc. for NASA.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.P. (renee.prange@obspm.fr).

Optically programmable electron spin memory using semiconductor quantum dots

Miro Kroutvar, Yann Ducommun, Dominik Heiss, Max Bichler, Dieter Schuh, Gerhard Abstreiter & Jonathan J. Finley

Walter Schottky Institut, Technische Universität München, Am Coulombwall 3, D-85748 Garching, Germany

The spin of a single electron subject to a static magnetic field provides a natural two-level system that is suitable for use as a quantum bit, the fundamental logical unit in a quantum computer^{1–3}. Semiconductor quantum dots fabricated by strain driven self-assembly⁴ are particularly attractive for the realization of spin quantum bits, as they can be controllably positioned⁵, electronically coupled⁶ and embedded into active devices^{7–10}. It has been predicted that the atomic-like electronic structure⁴ of such quantum dots suppresses coupling of the spin to the solid-state quantum dot environment^{11–14}, thus protecting the 'spin' quantum information against decoherence^{15,16}. Here we demonstrate a single electron spin memory device in which the electron spin can be programmed by frequency selective optical excitation. We use the device to prepare single electron spins in semiconductor quantum dots with a well defined orientation, and directly measure the intrinsic spin flip time and its dependence on magnetic field. A very long spin lifetime is obtained, with a lower limit of about 20 milliseconds at a magnetic field of 4 tesla and at 1 kelvin.

We begin by summarizing the operating principles of our optical spin storage device before presenting the measurements of the spin flip time, its dependence on magnetic field and the determination of the underlying mechanism. The structure of the devices investigated and the measurement techniques are summarized in Fig. 1. The samples consist of a single layer of self-assembled Ga(In)As quantum dots (QDs) embedded within the intrinsic region of a p-type GaAs Schottky photodiode¹⁷. Single electron–hole pair excitations (excitons) are generated directly in QD ground states by frequency-

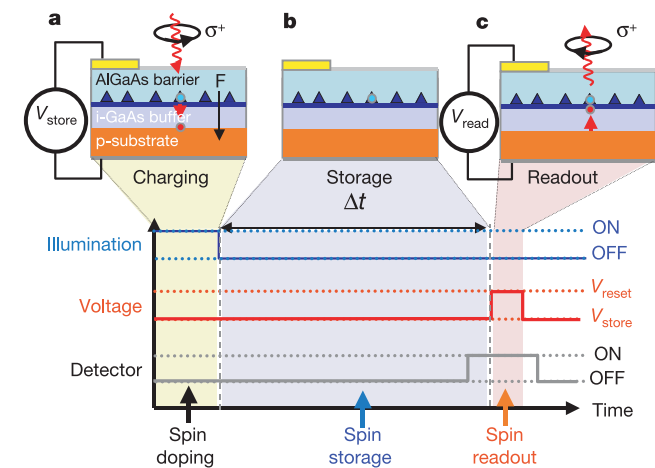


Figure 1 Schematic of devices and operation principle. **a**, Following resonant excitation into QD ground states, excitons are ionized owing to the axial electric field provided by the applied bias (V_{store}). The AlGaAs barrier above the QD layer inhibits electron escape from the QDs, ensuring photon-to-charge conversion. **b**, After charge separation, electrons are stored directly in the QDs where they were generated. **c**, Readout of the stored charge distribution by forward biasing the device ($V = V_{\text{read}}$). Holes drift back into the QDs, neutralizing the stored charge and generating a photon, the polarization of which probes the electron spin orientation—see Supplementary Information for further details.

selective optical excitation on the low-energy side of the QD ensemble. Photon-to-charge conversion is ensured by the electric field in the intrinsic region of the device ($\sim 160 \text{ kV cm}^{-1}$) orientated parallel to the QD growth axis. This results in holes tunnelling out of the QDs over timescales much faster than the exciton radiative lifetime⁸, while electrons remain stored owing to the presence of an asymmetric $\text{Ga}_{0.6}\text{Al}_{0.4}\text{As}$ barrier grown immediately above the QDs (Fig. 1a). Our previous measurements¹⁷ have confirmed that this technique enables selective generation of single electrons directly in the QD ground state where they remain over timescales of up to hours at low temperature (Fig. 1b)¹⁸. Generation of only one electron per dot is ensured by the few milli-electronvolt renormalization of the inter-band absorption energy following charging¹⁹.

Following a delay time Δt , the stored electron distribution is optically probed by forward biasing the Schottky junction, as depicted schematically in Fig. 1c. A drift current of holes then flows into the negatively charged dots and neutralizes the stored charge. The electron–hole pairs in the QDs then rapidly recombine over the radiative recombination time ($\approx 1 \text{ ns}$), generating a time-delayed electroluminescence (EL) signal that directly reflects the spectral distribution of stored charge a time Δt after generation. A single photon counter is gated ON immediately before the reset voltage pulse to detect the emitted photons and OFF again after $\tau_{\text{det}} \approx 500 \text{ ns}$ (Fig. 1), after which the exciton population has decayed completely. Further details of the device operation and experimental techniques can be found in Supplementary Information.

To extend the concepts introduced above from photon-to-charge conversion to optically induced electron spin orientation, a key consideration is the spin structure of the QD ground state exciton. The QD exciton states are constructed from electron (e) and heavy-hole (h) single-particle basis states with spin projections along the QD growth axis (z) of $J_{e,z} = +1/2\hbar$ or $-1/2\hbar$ ($e\uparrow$ or $e\downarrow$) and $J_{h,z} = -3/2\hbar, +3/2\hbar$ ($h\downarrow$ and $h\uparrow$) respectively²⁰. As a circularly polarized photon conveys one unit of angular momentum ($+1\hbar$ for σ^+ and $-1\hbar$ for σ^-) and the optical transition takes place to the crystal ground state, only the exciton states with $J_z = J_{e,z} + J_{h,z} = \pm 1\hbar$ ($e\downarrow h\uparrow$ and $e\uparrow h\downarrow$) are optically active. However, the $e\uparrow h\downarrow$

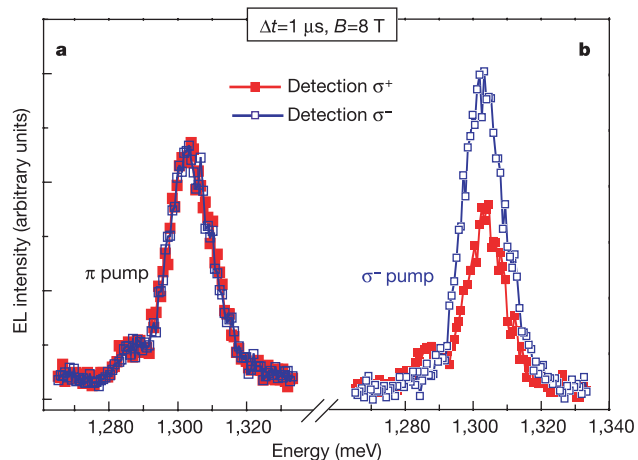


Figure 2 Spin storage spectra for $T = 10 \text{ K}$, high magnetic field $B = 8 \text{ T}$ and a short storage time of $\Delta t = 1 \mu\text{s}$. **a**, After randomly linearly polarized excitation, EL intensities for σ^+ (red, filled squares) and σ^- (blue, open squares) are equal. Both spin orientations are created equally. No thermalization into the lower Zeeman level occurs. **b**, Following σ^- excitation, co-polarized σ^- EL intensity (blue, open squares) is enhanced versus the counter-polarized σ^+ light (red, filled squares). Selective pumping of one Zeeman level exhibits a long-term spin memory: an approximately 65% degree of co-polarization for the data presented, eventually becoming as large as $\sim 85\%$ at $T = 1 \text{ K}$.

and $e\downarrow h\uparrow$ eigenstates are often mixed in dots with reduced symmetry, forming two linearly polarized eigenstates separated by the anisotropic e–h exchange splitting of a few times $10 \mu\text{eV}$ (δ_1) (refs 20, 21). In order to enable optical selection of pure spin states, all the measurements described below were performed with large static magnetic fields ($B \geq 4 \text{ T}$) applied parallel to the growth axis of the dots. This results in a Zeeman splitting $\Delta E_Z^{\text{ex}} = g^* \mu_B B$ of the QD exciton levels, where g^* is the excitonic g-factor. This transforms them into pure $e\uparrow h\downarrow$ and $e\downarrow h\uparrow$ eigenstates, providing that $\Delta E_Z^{\text{ex}} > \delta_1$ (ref. 20). Optical excitation using circularly polarized light with σ^- or σ^+ helicity then selectively generates electrons with up ($e\uparrow$) or down ($e\downarrow$) spin orientation, respectively. When the spin is read out after a time delay Δt , the degree of circular polarization of the emitted EL intensity (I), $P = (I^{\sigma^+} - I^{\sigma^-}) / (I^{\sigma^+} + I^{\sigma^-})$, then provides a direct probe of the electron spin orientation a time Δt after generation.

In order to verify the concepts introduced above, we performed spin storage measurements for a short delay time ($\Delta t = 1 \mu\text{s}$) and large magnetic field ($B = 8 \text{ T}$), while varying the polarization state of the excitation source. Figure 2 compares typical storage spectra recorded at $T = 10 \text{ K}$ following excitation with randomly linearly polarized (left panel) or σ^- circularly polarized light (right panel). By comparing the frequency and spatially selective nature of the excitation process, we estimate that $\sim 10,000$ dots are addressed by the measurement (see Supplementary Information). Following linear excitation, the storage spectra detected with σ^- and σ^+ polarization are equally intense because both spin orientations ($e\uparrow$ and $e\downarrow$) are generated with equal probability in the QDs. The lack of any circular polarization excludes the possibility that spin alignment due to inter Zeeman level thermalization occurs over such short storage times. In contrast, following σ^- excitation to generate $e\uparrow$ electrons in the lower Zeeman level, the storage EL is found to be strongly co-polarized (Fig. 2, right panel). At 10 K and $B = 8 \text{ T}$, the measured degree of circular polarization following σ^- excitation ($P(\sigma^-)$) was as large as $\sim 65\%$, but was found to increase strongly (up to $\sim 85\%$) with reducing temperature as discussed below. Repeating this measurement using σ^+ excitation to generate a population of $e\downarrow$ electrons in the upper Zeeman level again produced co-polarized emission, but with precisely the

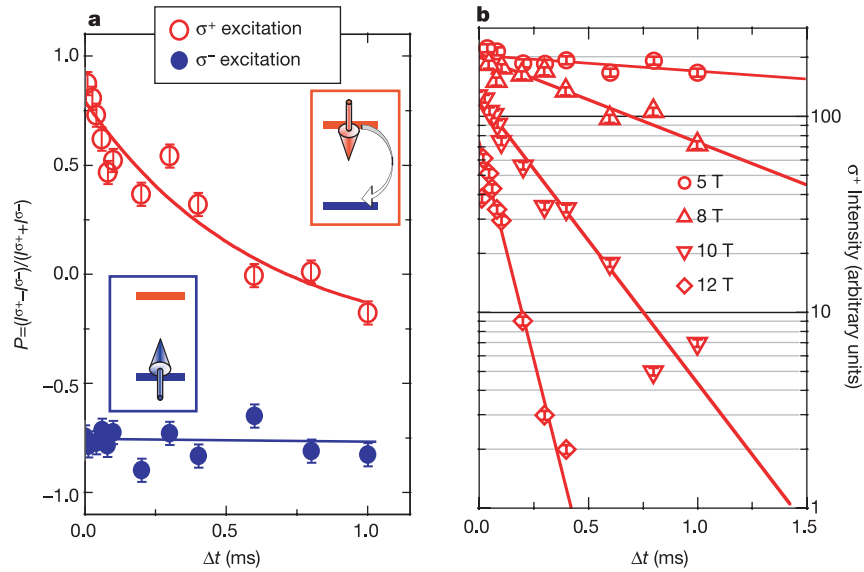


Figure 3 Electron spin dynamics at $T = 1$ K and $B = 8$ T for storage times up to 1 ms. **a**, The degree of circular polarization (P) at peak of the storage EL. Red: pumping electron spins into the upper Zeeman level. Temporal decay observed from co-polarized emission ($P > 0$) at short delay times to counter-polarized emission ($P < 0$) after $\Delta t > 0.75$ ms. Blue: pumping electrons into the lower Zeeman level results in co-polarized emission

($P < 0$) but no detectable dynamics. **b**, Semi-logarithmic plot of the temporal evolution of the co-polarized EL intensity following pumping electrons into the upper Zeeman level for different magnetic fields. T_1 and the related standard error is extracted from a least-squares mono-exponential fit of the observed decay transients.

opposite degree of circular polarization, $P(\sigma^+) \approx +65\%$. These results demonstrate that the storage signal exhibits a pronounced polarization memory for both spin orientations, arising from the reversible transfer from optical polarization to electron spin orientation (Fig. 1a) followed by spin storage for a time Δt (Fig. 1b) and back-transfer from electron spin orientation into optical polarization (Fig. 1c).

Following demonstration of this spin memory effect, we now consider the temporal evolution of $P(\sigma^+)$ and $P(\sigma^-)$, a measurement that reflects the electron spin dynamics in the QDs during the storage time (Δt). Figure 3a shows the degree of circular polarization of the storage EL, monitored at the peak position as a function of storage time in the range $0.001 < \Delta t$ (ms) < 1 at $B = 8$ T and a reduced lattice temperature of $T = 1$ K. $P(\sigma^-)$ was found to be as large as -85% with no detectable temporal evolution up to 1 ms (Fig. 3a, lower curve). In contrast, a very marked time evolution of $P(\sigma^+)$ is observed (Fig. 3a, upper curve). For $\Delta t \approx 0.001$ ms, $P(\sigma^+)$ is as large as $+85\%$, decaying over time to $\sim -10\%$ at $\Delta t \approx 1$ ms, as electrons relax from the upper to the lower Zeeman levels. At thermal equilibrium, the population ratio of the upper and lower Zeeman levels is $N_1/N_1 = \exp(-g_s^* \mu_B B / k_B T)$, indicating that $N_1/N_1 < 0.02$ under the present experimental conditions ($g_s^* \mu_B B \approx 0.36$ meV and $k_B T \approx 0.09$ meV). Therefore, the contrast of the observed decay from the upper level (N_1) is much larger, while spin flip mechanisms for $e \downarrow \Rightarrow e \uparrow$ and $e \uparrow \Rightarrow e \downarrow$ remain the same. To confirm this interpretation, we reversed the orientation of the magnetic field to invert the energetic ordering of the Zeeman levels. In this case, a strong characteristic temporal dependence of $P(\sigma^-)$ was observed as expected, while no measurable dynamics could be observed for $P(\sigma^+)$.

As the σ^- polarized signal from the lower Zeeman level exhibits no detectable dynamics over the timescales investigated, the temporal evolution of the σ^+ polarized EL intensity following σ^+ excitation (that is, $I^{\sigma^+}(t)$) was used to provide a direct measurement of the electron spin relaxation time (T_1) in QDs. Time resolved traces of $I^{\sigma^+}(t)$ in the range $\Delta t = 0.001$ –1 ms are presented in Fig. 3b for selected magnetic fields up to $B = 12$ T. For each magnetic field investigated, $I^{\sigma^+}(t)$ exhibits a mono-exponential decay,

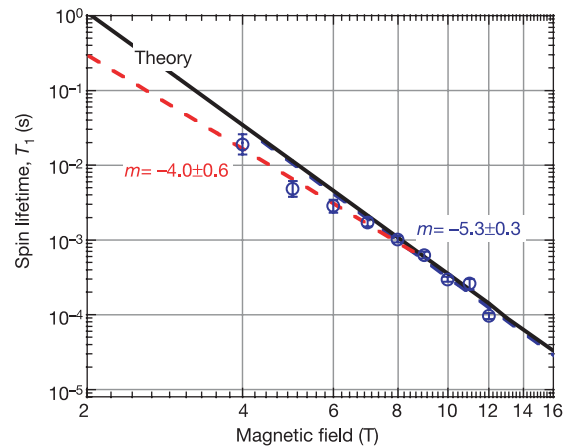


Figure 4 Double logarithmic plot of the spin lifetime T_1 versus the magnetic field at $T = 1$ K. T_1 increases from 0.1 ± 0.01 ms at $B = 12$ T to 20 ± 6 ms at $B = 4$ T (open circles). A least-squares fit of the B -field dependence for $B \geq 7$ T (blue dashed line) reveals a power-law dependence with an exponent $m = -5.3 \pm 0.3$, suggesting that inter Zeeman level spin scattering is due to spin-orbit coupling mediated by single piezoelectric phonons as discussed in the text. Solid line: at high B -fields, the theory of refs 13 and 14 accounts remarkably well for the measured data using reasonable parameters for the investigated Ga(In)As QDs ($|g_e| = 0.8$, $\hbar\omega_0 = 30$ meV). For $B < 7$ T, a softening of the exponent to $m = -4.0 \pm 0.6$ (red dashed line) is observed owing to the increasing phonon population.

$I^{\sigma^+}(t) = I_0^{\sigma^+} \exp(-t/T_1)$, as shown by the fits in Fig. 3b. For $B < 5$ T, storage times up to ~ 5 ms were measured to obtain reliable statistics. The decay time constants extracted from the data in Fig. 3b are found to be strongly dependent on the magnetic field, reducing markedly from $T_1 = 20 \pm 6$ ms to 0.1 ± 0.01 ms as the magnetic field increases from $B = 4$ to 12 T. The dependence of T_1 on the magnetic field is summarized in Fig. 4 on a double logarithmic representation. These data suggest a clear power-law dependence, $T_1 \propto B^m$, the best fit to the experimental data yielding a very large

exponent close to $m \approx -5$. We now consider the origin of this strong magnetic field dependence.

Spin-flip transitions between Zeeman levels must naturally proceed with conservation of energy, which can be ensured via the participation of acoustic phonons with an energy equal to the Zeeman energy, $E_{\text{ph}} = g_e \mu_B B$ (refs 13, 14). In this case, one-phonon scattering processes can mediate the spin-flip process by mixing of the Zeeman levels via the spin-orbit interaction^{13,14}. The spin relaxation time due to such a one-phonon admixture process in QDs has been calculated using perturbation theory^{13,14} to be:

$$\frac{1}{T_1} = A \frac{(g_e \mu_B B)^5}{\hbar (\hbar \omega_0)^4}$$

where A is a dimensionless constant that reflects the effective spin piezoelectric phonon coupling strength in the heterostructure, and $\hbar \omega_0$ is the QD single-particle level spacing. Using typical material properties applicable to GaAs (refs 13, 14) gives $A = 0.014$, which together with reasonable values for $|g_e| = 0.8$ (ref. 20) and $\hbar \omega_0 = 30 \text{ meV}$ (ref. 22) provide good quantitative agreement with our experimental data, as indicated by the full line in Fig. 4. The experimentally observed exponent $m = -4.9 \pm 0.2$ (not shown) is very close to the characteristic $T_1 \propto B^{-5}$ dependence expected for such one-phonon scattering processes, providing (to our knowledge) the first experimental evidence that inter Zeeman level spin-flip transitions in QDs are dominated by one-phonon scattering processes at low temperature. Closer inspection of the data presented in Fig. 4 indicates a softening of the exponent with reducing magnetic field. As shown by the blue and red dashed lines in Fig. 4, the best fit to the data for higher magnetic fields is very well described by a single exponent $m = -5.3 \pm 0.3$ (indicative of one-phonon processes, as discussed above) with the data at lower field better described by a weaker exponent $m = -4.0 \pm 0.6$. Such behaviour is expected for one-phonon processes^{13,14}, reflecting a transition between a low-temperature regime ($k_B T < g_e \mu_B B$), with the $T_1 \propto B^{-5}$ dependency discussed above, to a $T_1 \propto (k_B T)^{-1} B^{-4}$ dependency when $k_B T$ becomes comparable to $g_e \mu_B B$. In our experiment, the Zeeman splitting is comparable to the thermal energy for $B \approx 2\text{--}4 \text{ T}$, in good accord with the observed softening of the exponent at low magnetic fields.

The longest T_1 time that we measured ($\sim 20 \text{ ms}$ at $T = 1 \text{ K}$ and $B = 4 \text{ T}$) is limited by the sensitivity of our detection scheme, and represents a lower limit. The absence of any observable saturation of the T_1 in Fig. 4 suggests that the relaxation time should be much longer at lower fields. For example, extrapolating the observed $T_1 \propto B^{-4}$ dependency to lower fields would indicate $T_1 = 80 \text{ ms}$ at $\sim 3 \text{ T}$, reaching $\sim 1 \text{ s}$ at $B \approx 1.6 \text{ T}$. However, additional spin relaxation mechanisms (such as that due to hyperfine coupling to the nuclear spin²³) will potentially limit the intrinsic T_1 times that are accessible as B approaches zero. Recent theoretical work has suggested that if spin-orbit coupling mediated by single phonons is the dominant spin scattering mechanism, then the spin coherence time should approach the limit $T_2 = 2T_1$. Our results demonstrate that this mechanism does indeed dominate, suggesting that the spin coherence times for such optically generated electrons spins may also be long¹⁵. □

Received 9 July; accepted 10 September 2004; doi:10.1038/nature03008.

1. Loss, D. & DiVincenzo, D. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
2. Kane, B. E. A silicon-based nuclear spin quantum computer. *Nature* **393**, 119–120 (1998).
3. Elzerman, J. M. *et al.* Single-shot read-out of an individual electron spin in a quantum dot. *Nature* **430**, 431–435 (2004).
4. Arakawa, Y. & Tarucha, S. (eds) *Proc. 2nd Int. Conf. on Semiconductor Quantum Dots (QD 2002)*, *Physica Status Solidi B* **238**, 229–372, (2003).
5. Heidemeyer, H., Denker, U., Müller, C. & Schmidt, O. G. Morphology response to strain field interferences in stacks of highly ordered quantum dot arrays. *Phys. Rev. Lett.* **91**, 196103 (2003).
6. Bayer, M. *et al.* Coupling and entangling of quantum states in quantum dot molecules. *Science* **291**, 451–453 (2001).
7. Zrenner, A. *et al.* Coherent properties of a two-level system based on a quantum-dot photodiode. *Nature* **418**, 612–614 (2002).

8. Oulton, R. *et al.* Manipulation of the homogeneous linewidth of an individual In(Ga)As quantum dot. *Phys. Rev. B* **66**, 045313 (2002).
9. Hanson, R., Witkamp, B., Vandersypen, L. M. K., Willems van Beveren, L. H. & Elzerman, J. M. Zeeman energy and spin relaxation in a one-electron quantum dot. *Phys. Rev. Lett.* **91**, 196802 (2002).
10. Fujisawa, T., Austing, D. G., Tokura, Y., Hirayama, Y. & Tarucha, S. Nonequilibrium transport through a vertical quantum dot in the absence of spin-flip energy relaxation. *Phys. Rev. Lett.* **88**, 236802 (2002); Allowed and forbidden transitions in artificial hydrogen and helium atoms. *Nature* **419**, 278 (2002).
11. Meier, F. & Zakharchenya, B. (eds) *Modern Problems in Condensed Matter Sciences Vol. 8, Optical Orientation* (North Holland, Amsterdam, 1984).
12. Khaetskii, A. V. & Nazarov, Y. V. Spin relaxation in semiconductor quantum dots. *Phys. Rev. B* **61**, 12639–12642 (2000).
13. Khaetskii, A. V. & Nazarov, Y. V. Spin flip transitions between Zeeman sublevels in semiconductor quantum dots. *Phys. Rev. B* **64**, 125316 (2001).
14. Woods, L. M., Reinecke, T. L. & Lyanda-Geller, Y. Spin relaxation in quantum dots. *Phys. Rev. B* **66**, 161318 (2002).
15. Gupta, J. A., Awschalom, D. D., Peng, X. & Alivisatos, A. P. Spin coherence in semiconductor quantum dots. *Phys. Rev. B* **59**, R10421–R10424 (1999).
16. Golovach, V. N., Khaetskii, A. V. & Loss, D. Phonon induced decay of the electron spin in quantum dots. *Phys. Rev. Lett.* **93**, 016601 (2004).
17. Kroutvar, M. *et al.* Wavelength selective charge storage in self-assembled InGaAs/GaAs quantum dots. *Appl. Phys. Lett.* **83**, 443–445 (2003).
18. Finley, J. J. *et al.* Electrical detection of optically induced charge storage in self-assembled InAs quantum dots. *Appl. Phys. Lett.* **73**, 2618–2620 (1998).
19. Finley, J. J. *et al.* Charged and neutral exciton complexes in individual self-assembled In(Ga)As quantum dots. *Phys. Rev. B* **63**, 073307 (2001).
20. Bayer, M. *et al.* Fine structure of neutral and charged excitons in self-assembled In(Ga)As/(Al)GaAs quantum dots. *Phys. Rev. B* **65**, 195315 (2002).
21. Langbein, W. *et al.* Control of fine-structure splitting and biexciton binding in $\text{In}_x\text{Ga}_{1-x}\text{As}$ quantum dots by annealing. *Phys. Rev. B* **69**, 161301 (2004).
22. Oulton, R. *et al.* Continuum transitions and phonon coupling in single self-assembled Stranski-Krastanow quantum dots. *Phys. Rev. B* **68**, 235301 (2003).
23. Khaetskii, A. V., Loss, D. & Glazman, L. Electron spin evolution induced by interaction with nuclei in a quantum dot. *Phys. Rev. B* **67**, 195329 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank V. Golovach and D. Loss for discussions. We also thank the DFG for financial support and Attocube GmbH for technical support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.J.F. (finley@wsi.tum.de).

Lead-free piezoceramics

Yasuyoshi Saito¹, Hisaaki Takao¹, Toshihiko Tani¹,
Tatsuhiko Nonoyama², Kazumasa Takatori¹, Takahiko Homma¹,
Toshiatsu Nagaya² & Masaya Nakamura²

¹Toyota Central R&D Laboratories, Inc., Nagakute, Aichi, 480-1192, Japan

²DENSO Corporation, 1-1, Showa-cho, Kariya, Aichi, 448-8861, Japan

Lead has recently been expelled from many commercial applications and materials (for example, from solder, glass and pottery glaze) owing to concerns regarding its toxicity. Lead zirconium titanate (PZT) ceramics are high-performance piezoelectric materials, which are widely used in sensors, actuators and other electronic devices; they contain more than 60 weight per cent lead. Although there has been a concerted effort to develop lead-free piezoelectric ceramics, no effective alternative to PZT has yet been found^{1–14}. Here we report a lead-free piezoelectric ceramic with an electric-field-induced strain comparable to typical actuator-grade PZT. We achieved this through the combination of the discovery of a morphotropic phase boundary in an alkaline niobate-based perovskite solid solution, and the development of a processing route leading to highly (001) textured polycrystals. The ceramic exhibits a piezoelectric constant d_{33} (the induced charge per unit force applied in the same

direction) of above 300 picocoulombs per newton (pC N^{-1}), and texturing the material leads to a peak d_{33} of 416 pC N^{-1} . The textured material also exhibits temperature-independent field-induced strain characteristics.

The piezoelectric properties of PZT materials can be improved by the formation of a morphotropic phase boundary (MPB) between the tetragonal and rhombohedral phases in solid solutions of perovskite-type PbTiO_3 and PbZrO_3 (refs 15–17). For the development of new, lead-free piezoelectric materials, we designed a compositional formation of MPB between a different pair of crystal structures—namely, the pseudo-ilmenite-type and perovskite-type structures, having rather different lattice forms and unit sizes from each other. We anticipated the formation of a new MPB in the perovskite-rich region by the dissolution of a small amount of the pseudo-ilmenite-structured material, causing a lattice distortion for the structural phase transition. The general need for stable piezoelectric characteristics over a wide temperature range made us select high Curie-temperature ($T_C > 250^\circ\text{C}$) end members: orthorhombic perovskite-type $(\text{K}_{0.5}\text{Na}_{0.5})\text{NbO}_3$ ($T_C = 415^\circ\text{C}$) and hexagonal pseudo-ilmenite-type LiTaO_3 ($T_C = 615^\circ\text{C}$). Besides the formation of a MPB, we also exploited the hybridization of covalency onto ionic bonding for further improvement in piezoelectricity, on the basis of Cohen's calculation for the titanate-perovskite system¹⁸. In addition to LiTaO_3 , we also used LiSbO_3 as an end member for the compositional study with $(\text{K}_{0.5}\text{Na}_{0.5})\text{NbO}_3$, because the higher electronegativities of Sb and Ta compared to Nb were expected to make the alkaline niobate-based perovskite more covalent.

Figure 1a shows the result of the MPB study: a phase diagram determined by X-ray diffraction (XRD) measurements at room temperature (25°C) for conventionally sintered and unpoled ceramic specimens, and d_{31} (the induced charge per unit force applied in the perpendicular direction) values for the specimens poled at 5 kV mm^{-1} . ('Poled' indicates that the specimen has been given a macroscopic polar axis by altering the direction of the spontaneous polarization with an applied high electric field.) The compositions with high d_{31} values, shown in Fig. 1 as LF1, LF2 and LF3, are found in a tetragonal phase area near the MPB formed between orthorhombic and tetragonal phases in the alkaline niobate-based perovskite system.

The Curie temperature is controllable between 170 and 500°C in this compositional range, such that additions of Li and Ta elements shift the Curie temperature higher and lower, respectively. The highest piezoelectric charge sensor d_{33} constant in the $(\text{K},\text{Na})\text{NbO}_3$ – LiTaO_3 system was found to be 230 pC N^{-1} with a Curie temperature of 323°C at the LF3 composition. The introduction of the solid solution in the $(\text{K},\text{Na})\text{NbO}_3$ – LiTaO_3 – LiSbO_3 pseudo-ternary system further enhanced the piezoelectric performance, including a d_{33} constant up to as high as 300 pC N^{-1} , with a Curie temperature of 253°C (at the LF4 composition, which is also a tetragonal phase at 25°C in the unpoled state; Fig. 1b). These excellent piezoelectric properties exceed the d_{33} constant (200 pC N^{-1}) of the non-doped PZT (PZT1 in Fig. 1b).

By additional engineering of the microstructural design, we developed a novel processing route for producing textured polycrystals of the alkaline niobate-based compositions, LF3 and LF4. The d_{33} constants and Curie temperatures are shown in Fig. 1b for the textured ceramics LF3T and LF4T, for which the Lotgering's factors¹⁹ of the $\langle 001 \rangle$ orientation are 92% and 91%, respectively. The enhanced d_{33} constants are 373 and 416 pC N^{-1} for LF3T and LF4T, respectively, which are 1.8 and 1.6 times as large as those of non-textured ceramics, LF3 and LF4, respectively.

For the successful fabrication of textured ceramics with high d_{33} values, one of the most important technologies is the preparation of plate-like particles, which are used as a template. The hot-working (hot-forging²⁰ and hot-pressing²¹) method cannot give a texture to such a pseudo-isotropic system, and no proper template materials have been available for the templated grain growth (TGG) and reactive-templated grain growth (RTGG)^{22–27} of the alkaline niobate-based perovskite. We have developed a new synthesis technique for plate-like NaNbO_3 particles as templates for $\langle 001 \rangle$ oriented $(\text{K},\text{Na})\text{NbO}_3$ -based ceramics. This process comprises three steps as follows: (1) Bismuth layer-structured plate-like $\text{Bi}_{2.5}\text{Na}_{3.5}\text{Nb}_5\text{O}_{18}$ (BiNN5) particles (shown in Fig. 2b and c) are synthesized as a precursor by the molten-salt synthesis method. (2) Using this precursor, plate-like NaNbO_3 particles (Fig. 2d and e), identified by JCPDS powder diffraction file card No.33-1270, are synthesized through a topochemical reaction, in which NaNbO_3 is

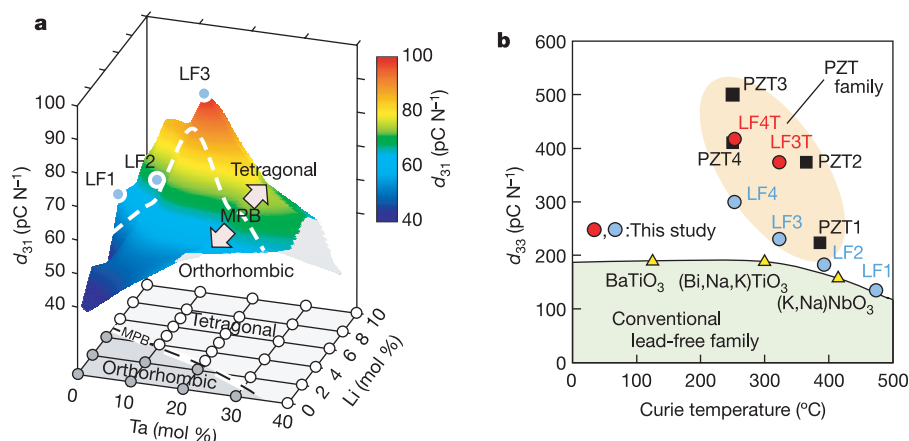


Figure 1 Piezoelectric sensor performances for the lead-free (LF) piezoelectric ceramics. **a**, Piezoelectric charge sensor d_{31} constants at 25°C as functions of Li and Ta contents for developed $(\text{K}_{0.5}\text{Na}_{0.5})_{1-x}\text{Li}_x(\text{Nb}_{1-y}\text{Ta}_y)\text{O}_3$ ceramics. The compositions of LF1, LF2 and LF3 are $(x,y) = (0.06,0)$, $(x,y) = (0.04,0.10)$ and $(x,y) = (0.03,0.20)$, respectively. Note the morphotropic phase boundary (MPB) between tetragonal and orthorhombic phases. The phase diagram was determined by XRD at 25°C for conventionally sintered and unpoled ceramic specimens; d_{31} values are shown for specimens poled at 5 kV mm^{-1} . **b**, Comparison of the piezoelectric charge sensor d_{33} constants at 25°C

among developed LF ceramics, previously reported LF ceramics (refs 1–3), and conventional PZT ceramics (refs 15, 17) as a function of Curie temperature. LF4: $(\text{K}_{0.44}\text{Na}_{0.52}\text{Li}_{0.04})(\text{Nb}_{0.86}\text{Ta}_{0.10}\text{Sb}_{0.04})\text{O}_3$. LF3T and LF4T: textured ceramics with the same compositions as LF3 and LF4, respectively. PZT1: $\text{Pb}(\text{Zr}_{0.52}\text{Ti}_{0.48})\text{O}_3$. PZT2: $\text{Pb}_{0.988}(\text{Zr}_{0.48}\text{Ti}_{0.52})_{0.976}\text{Nb}_{0.024}\text{O}_3$. PZT3: commercially available PZT. PZT4: $(\text{Pb}_{0.85}\text{Ba}_{0.15})_{0.9925}\text{La}_{0.005}(\text{Zr}_{0.52}\text{Ti}_{0.48})\text{O}_3$. The d_{33} values were measured for specimens poled at 5 kV mm^{-1} .

formed by ion exchange of Na for Bi ions on BiNN5 with preserved particle morphology and a developed {001} plane of the perovskite (Fig. 2a). (3) This NaNbO_3 platelet is used as a reactive template, and textured $(\text{K,Na})\text{NbO}_3\text{--LiTaO}_3\text{--LiSbO}_3$ polycrystals are synthesized by the RTGG method.

Figure 3a and b shows the scanning electron micrograph (SEM) image and X-ray diffraction (XRD) profile of textured $(\text{K}_{0.44}\text{Na}_{0.52}\text{Li}_{0.04})(\text{Nb}_{0.84}\text{Ta}_{0.10}\text{Sb}_{0.06})\text{O}_3$ ceramic (LF4T) in comparison to those of the non-textured ceramic (LF4) with the same composition. It should be noted that the textured ceramic gives brick-layer-like quadrangular grains, which align parallel to the tape-casting plane, and the {001} diffraction peaks of the textured ceramic in the XRD pattern are clearly high compared to those of the non-textured ceramic. In addition, the XRD patterns show that the phases of the textured and non-textured ceramics in the unpoled state are tetragonal at 25 °C. From this result, it is considered that the transformation from a low-temperature orthorhombic phase to a high-temperature tetragonal phase occurs below 25 °C.

In order to examine the applicability of the developed materials for high-power devices, actuator performances were evaluated. The electric-field-induced strain was measured for the textured LF4T, non-textured LF4 and the conventional PZT ceramic (PZT4) for actuator application under a high electric field from 0 to $2,000\text{ V mm}^{-1}$ with a triangular wave^{28–30}.

For the non-textured LF4 (Fig. 4a), the electric-field-induced strain was unstable within this temperature range, and had an anomaly: the highest value occurred at 40 °C. It is considered that

this anomaly is related to the electric-field-induced tetragonal-to-orthorhombic phase transformation, shifting the phase transformation temperature from below 25 °C in the unpoled specimen to 40 °C under a high electric field. In order to clarify the origin of this

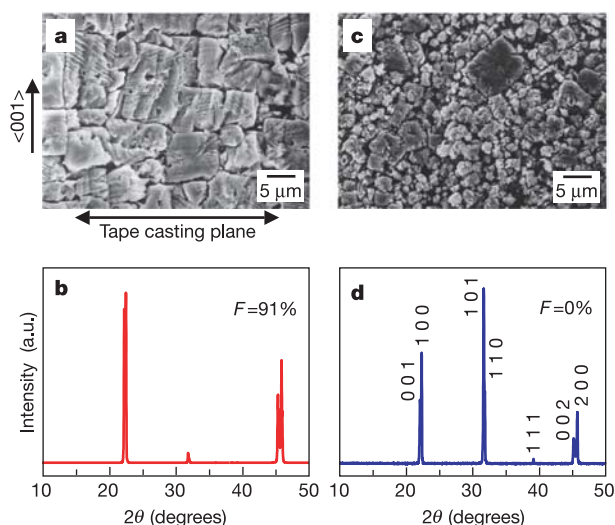


Figure 3 SEM images of etched cross-sections and X-ray diffraction profiles of textured and non-textured ceramics. **a, b**, {001} oriented ceramic (LF4T). The SEM image shows a cross-section perpendicular to the casting plane. **c, d**, Non-textured ceramic (LF4) with the same composition as LF4T. *F* denotes the {001} axis orientation factor evaluated by Lotgering's method.

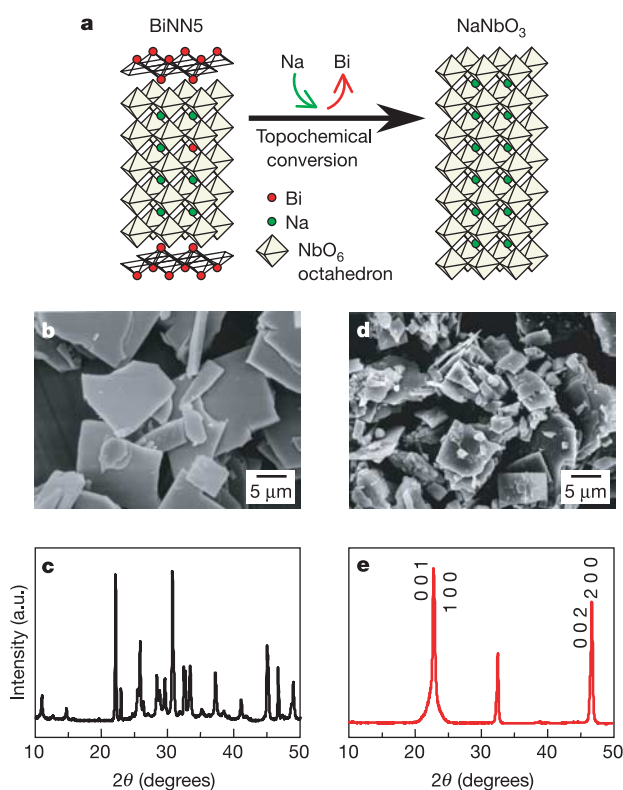
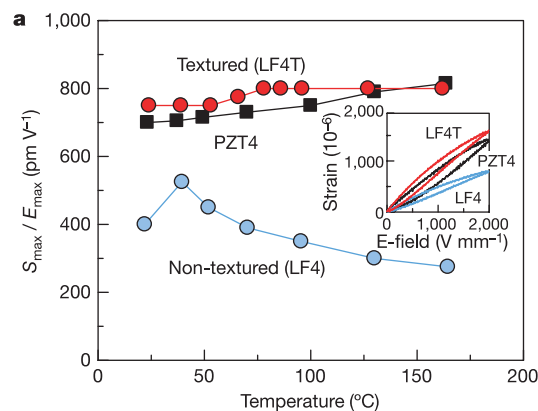


Figure 2 Schematic diagram of topochemical conversion from bismuth layer-structured BiNN5 particles to plate-like NaNbO_3 particles. **a**, Crystal structures of plate-like BiNN5 and NaNbO_3 particles. **b, c**, SEM image and X-ray diffraction profile of plate-like BiNN5 particles used as precursor. **d, e**, SEM image and X-ray diffraction profile of plate-like NaNbO_3 particles. X-ray diffraction profile of NaNbO_3 particles is characterized by pseudo-tetragonal Miller indices. BiNN5 particles are completely converted into the regular perovskite-structured NaNbO_3 particles with a preserved plate-like shape.



b

Piezoelectric property		LF4T	PZT4
Curie temperature	T_C (°C)	253	250
Piezoelectric coupling constant	K_p	0.61	0.60
Piezoelectric charge sensor constant	d_{31} (pC N ⁻¹)	152	170
	d_{33} (pC N ⁻¹)	416	410
Piezoelectric voltage constant	g_{31} (10 ⁻³ V m N ⁻¹)	11.0	8.3
	g_{33} (10 ⁻³ V m N ⁻¹)	29.9	20.2
Dielectric constant	$\epsilon_{33}^T / \epsilon_0$	1,570	2,300
Normalized strain	$S_{\max} / E_{\max}$ (pm V ⁻¹)	750	700

Figure 4 Actuator performances of the developed lead-free piezoelectric ceramics. **a**, Temperature dependences of electric-field-induced longitudinal strain for the textured (LF4T) and non-textured (LF4) ceramics. Inset, electric-field-induced strain curve for LF4T and LF4 at 25 °C. $S_{\max}$ and $E_{\max}$ denote the maximum strain and the maximum electric field strength, respectively. **b**, Piezoelectric properties of LF4T and PZT4. Dielectric constants were measured at 1 kHz.

anomaly, morphotropic phases need to be determined under high electric-field driving. However, this evaluation is very difficult to perform, and is left for future studies.

On the other hand, it is obvious in Fig. 4a that the textured LF4T, which is the $\langle 001 \rangle$ orientated ceramic, exhibited a large strain nearly double of that of the non-textured LF4 ceramic, and the strain was even larger than the field-induced strain for PZT4. In addition, we discovered that the texture given to LF4 not only enhanced the field-induced strain but also improved the temperature coefficient of normalized strain ($S_{\max}/E_{\max}$), as shown in Fig. 4a, where $S_{\max}$ and $E_{\max}$ denote the maximum strain and the maximum electric field strength, respectively. That is to say, the texture stabilizes the temperature-dependent strain characteristic of LF4T, even though its transformation temperature between orthorhombic and tetragonal phases is the same as that of non-textured LF4 in the unpoled state. Furthermore, it should be noted that the flat, temperature-independent characteristic of LF4T is even more prominent than PZT4, since the strain deviation value of 6.5% for LF4T between room temperature and 160 °C is smaller than that of 15% for PZT4. This piezoelectric strain behaviour of LF4T is of great importance for temperature-independent actuator devices.

The temperature-independent strain characteristics (for a given texture) are the result of changes in the amplitude of, and in the ratio between, two strain components; one from a lattice motion and the other from domain wall motion^{28–30}. Those two field-induced strain components must be dependent on temperature and microstructure (that is, textured or non-textured). In order to clarify the underlying mechanism, a measurement system needs to be developed for the separate evaluation of strain components from the two different origins.

The overall performance of the developed LF4T ceramic is listed in Fig. 4b with that of the typical high-performance PZT ceramic (PZT4). It is clear that most of the piezoelectric properties are comparable to those of the PZT. This high performance leads us to expect that the developed materials are leading candidates for environmentally friendly piezoelectric devices. □

Methods

$\text{Bi}_{2.5}\text{Na}_{3.5}\text{Nb}_5\text{O}_{18}$ (BiNN5) platelet was synthesized at 1,100 °C using molten NaCl salt as a flux. Using a topochemical reaction, NaNbO_3 platelet was synthesized from BiNN5 and a complementary reactant, Na_2CO_3 , in a NaCl flux at 950 °C, and Bi_2O_3 , the by-product, was removed. The synthesized NaNbO_3 had the same morphology as BiNN5 , 0.5 μm in thickness and 10–15 μm in side length of a developed area, and consisted of a single-phase perovskite with no secondary phase detected by XRD. The NaNbO_3 platelets (as reactive templates) and complementary reactants—equiaxed NaNbO_3 , KNbO_3 , KTaO_3 , LiSbO_3 and NaSbO_3 particles—were mixed, tape-cast and stacked. The textured ($\text{K}_{0.44}\text{Na}_{0.52}\text{Li}_{0.04}$)($\text{Nb}_{0.84}\text{Ta}_{0.10}\text{Sb}_{0.06}$) O_3 polycrystal (LF4T) was prepared by sintering the stacked tape at 1,135 °C.

The piezoelectric d_{31} constants were measured by the resonance anti-resonance method with an impedance analyser (Agilent, HP4194A). The piezoelectric d_{33} constants were measured by the direct piezoelectric method using a piezo- d_{33} meter (Institute of Acoustic Academia Sinica, model ZJ-4B). The degree of tetragonal $\langle 001 \rangle$ axis orientation, F , was evaluated by the Lotgering's equation using an X-ray diffractometer (Rigaku, RINT TTR2, $\text{CuK}\alpha$ radiation).

Received 25 May; accepted 3 September 2004; doi:10.1038/nature03028.
Published online 31 October 2004.

- Haertling, G. E. Ferroelectric ceramics: History and technology. *J. Am. Ceram. Soc.* **82**, 797–818 (1999).
- Takenaka, T. & Nagata, H. Present status of non-lead-based piezoelectric ceramics. *Key Eng. Mater.* **157–158**, 57–64 (1999).
- Jaeger, R. E. & Egerton, L. Hot pressing of potassium-sodium niobates. *J. Am. Ceram. Soc.* **45**, 209–213 (1962).
- Dungan, R. H. & Golding, R. D. Polarization of NaNbO_3 - KNbO_3 ceramic solid solutions. *J. Am. Ceram. Soc.* **48**, 601 (1965).
- Haertling, G. H. Properties of hot-pressed ferroelectric alkali niobate ceramics. *J. Am. Ceram. Soc.* **50**, 329–330 (1967).
- Egerton, L. & Bieling, C. A. Isostatically hot-pressed sodium-potassium niobate transducer material for ultrasonic devices. *Ceram. Bull.* **47**, 1151–1156 (1968).
- Aurivillius, B. Mixed bismuth oxides with layer lattices. *Ark. Kemi* **1**, 499 (1949).
- Wood, A. Polymorphism in potassium niobate, sodium niobate, and other ABO_3 compounds. *Acta Crystallogr.* **4**, 353–362 (1951).
- Buhrer, C. F. Some properties of bismuth perovskites. *J. Chem. Phys.* **36**, 798–803 (1962).
- Nitta, T. Properties of sodium-lithium niobate solid solution ceramics with small lithium concentrations. *J. Am. Ceram. Soc.* **51**, 626–629 (1968).

- Scot, B. A., Giess, E. A., Burns, G. & O'Kane, D. F. Alkali-rare earth niobates with the tungsten bronze-type structure. *Mater. Res. Bull.* **3**, 831–842 (1968).
- Hellwege, K.-H., Hellwege, A. M., Mitsui, T. & Nomura, S. (eds) *Numerical Data and Functional Relationships in Science and Technology*, New Series Group 3: *Crystal and Solid State Physics*, Vol. 16, *Ferroelectrics and Related Substances* Subvol. a, *Oxides* (Springer, Berlin, 1981).
- Mitsui, T. & Nakamura, E. (eds) *Numerical Data and Functional Relationships in Science and Technology*, New Series Group 3: *Crystal and Solid State Physics*, Vol. 28, Suppl. and Extension to Vol. 16, *Ferroelectrics and Related Substances* Subvol. a, *Oxides* (Springer, Berlin, 1990).
- Takenaka, T. & Nagata, H. *Program Summary and Extended Abstract of the 11th US-Japan Seminar on Dielectric and Piezoelectric Ceramics* 237–244 (US-Japan seminar committee, Sapporo, Hokkaido, Japan, 2003).
- Jaffe, B., Roth, R. S. & Marzullo, S. Properties of piezoelectric ceramics in the solid-solution series lead titanate zirconate-lead oxide: tin oxide and lead titanate-lead hafnate. *J. Res. Natl Bur. Stand.* **55**, 239–254 (1955).
- Zhang, S. J., Randall, C. A. & Shrout, T. R. High Curie temperature piezocrystals in the BiScO_3 - PbTiO_3 perovskite system. *Appl. Phys. Lett.* **83**, 3150–3152 (2003).
- Jaffe, B., Cook, W. R. & Jaffe, H. *Piezoelectric Ceramics* (Academic, New York, 1971).
- Cohen, R. E. Origin of ferroelectricity in perovskite oxides. *Nature* **358**, 136–138 (1992).
- Lotgering, F. K. Topotactical reactions with ferrimagnetic oxides having hexagonal crystal structures — I. *J. Inorg. Nucl. Chem.* **9**, 113–123 (1959).
- Takenaka, T. & Sakata, K. Grain orientation and electrical properties of hot-forged $\text{Bi}_4\text{Ti}_3\text{O}_{12}$ ceramics. *Jpn. J. Appl. Phys.* **19**, 31–39 (1980).
- Kimura, T., Yoshimoto, T., Iida, N., Fujita, Y. & Yamaguchi, T. Mechanism of grain orientation during hot-pressing of bismuth titanate. *J. Am. Ceram. Soc.* **72**, 85–89 (1989).
- Watanabe, H., Kimura, T. & Yamaguchi, T. Sintering of platelike bismuth titanate powder compacts with preferred orientation. *J. Am. Ceram. Soc.* **74**, 139–147 (1991).
- Brahmarout, B., Messing, G. L., Trolier-McKinstry, S. & Selvaraj, U. in *Proc. 10th IEEE Int. Symp. on Applications of Ferroelectrics* Vol. 2 (eds Kulwicki, B., Amin, A. & Safari, A.) 883–886 (Institute of Electrical and Electronic Engineers (IEEE), Piscataway, NJ, 1996).
- Horn, J. A., Zhang, S. C., Selvaraj, U., Messing, G. L. & Trolier-McKinstry, S. Templated grain growth of textured bismuth titanate. *J. Am. Ceram. Soc.* **82**, 921–926 (1999).
- Tani, T. Crystalline-oriented bulk ceramics with a perovskite-type structure. *J. Korean Phys. Soc.* **32** (Suppl. Iss.), S1217–S1220 (1998).
- Takeuchi, T., Tani, T. & Saito, Y. Piezoelectric properties of bismuth layer-structured ferroelectric ceramics with a preferred orientation processed by the reactive templated grain growth method. *Jpn. J. Appl. Phys.* **38**, 5553–5556 (1999).
- Sugawara, T., Shimizu, M., Kimura, T., Takatori, K. & Tani, T. Fabrication of grain oriented barium titanate. *Ceram. Trans.* **136**, 389–406 (2003).
- Saito, Y. Measurement of complex piezoelectric d_{33} constant in ferroelectric ceramics under high electric field driving. *Jpn. J. Appl. Phys.* **34**, 5313–5319 (1995).
- Saito, Y. Measurement system for electric field-induced strain by use of displacement magnification technique. *Jpn. J. Appl. Phys.* **35**, 5168–5173 (1996).
- Saito, Y. Hysteresis curve of X-ray diffraction peak intensity in lead zirconate titanate ceramics. *Jpn. J. Appl. Phys.* **36**, 5963–5969 (1997).

Acknowledgements We thank T. Takeuchi for technical contributions to the development of the processing; T. Saito for discussions; and N. Watanabe, M. Uoshima, H. Morisaka, Y. Aoki, K. Horibuchi, K. Hisazato, M. Okumura, M. Okano, K. Nomura and S. Tsuru for technical assistance.

Competing interests statement The authors declare competing financial interests: details accompany the paper on Nature's website (<http://www.nature.com/nature>).

Correspondence and requests for materials should be addressed to Y.S. (ysaito@mosk.tytlabs.co.jp).

Fire-induced erosion and millennial-scale climate change in northern ponderosa pine forests

Jennifer L. Pierce¹, Grant A. Meyer¹ & A. J. Timothy Jull²

¹Department of Earth and Planetary Science, University of New Mexico, Albuquerque, New Mexico 87131, USA

²NSF-Arizona AMS Facility, The University of Arizona, Tucson, Arizona 85721, USA

Western US ponderosa pine forests have recently suffered extensive stand-replacing fires followed by hillslope erosion and sedimentation^{1–4}. These fires are usually attributed to increased stand density as a result of fire suppression, grazing and other land use, and are often considered uncharacteristic or

unprecedented^{1–3}. Tree-ring records from the past 500 years indicate that before Euro-American settlement, frequent, low-severity fires maintained open stands^{1–3}. However, the pre-settlement period between about AD 1500 and AD 1900 was also generally colder than present^{5–10}, raising the possibility that rapid twentieth-century warming promoted recent catastrophic fires. Here we date fire-related sediment deposits in alluvial fans in central Idaho to reconstruct Holocene fire history in xeric ponderosa pine forests and examine links to climate. We find that colder periods experienced frequent low-severity fires, probably fuelled by increased understory growth. Warmer periods experienced severe droughts, stand-replacing fires and large debris-flow events that comprise a large component of long-term erosion¹¹ and coincide with similar events in sub-alpine forests of Yellowstone National Park¹². Our results suggest that given the powerful influence of climate, restoration of processes typical of pre-settlement times may be difficult in a warmer future that promotes severe fires.

Fire regimes in western US conifer forests vary spatially with local climate^{2,13,14}. Cool, moist, high-elevation spruce-fir forests experience severe, infrequent stand-replacing crown fires, whereas frequent light surface fires are considered typical of warm, xeric ponderosa pine (*Pinus ponderosa*) forests^{1–3,13}. Over the twentieth century, fire size and severity have increased in most ponderosa pine forests, including those in the northern Rocky Mountains. From 1908 to 2000, canopy fires burned 50% of the Boise National Forest in central Idaho, mostly during severe droughts in 1926–35 and 1986–2000. Recent stand-replacing fires in this area have led to numerous large debris flows and sediment-charged floods in steep mountain drainages⁴. Sediment yields from single large sedimentation events following severe 1989 and 1994 fires⁴ are three orders of magnitude greater than annual sediment yields measured without large events¹¹.

To assess the role of longer-term climatic variations on fire and erosion in northern ponderosa pine forests, we date fire-related deposits in tributary alluvial fans of the South Fork Payette River area, central Idaho (SFP; Fig. 1). These fans receive sediment from small (0.01–6 km²), steep basins in weathered Idaho batholith granitic rocks, where post-fire erosion is likely. Annual precipitation is ~600–1,000 mm, with a marked summer drought conducive to fires. Ponderosa pines are common at lower elevations (~850–1,300 m) and on dry south-facing slopes. Mixed Douglas-fir (*Pseudotsuga menziesii*) and ponderosa pine stands dominate mid-elevation (~1,500–1,800 m) and north-facing slopes, and mixed conifers dominated by Douglas-fir and

Engelmann spruce (*Picea engelmannii*) characterize high elevations around 2,000 m.

Post-fire debris-flow events and sediment-charged floods are produced via two primary mechanisms in the SFP area⁴. (1) After severe burns, reduced infiltration and smooth soil surfaces greatly increase surface runoff, typically during moderate- to high-intensity convective storms¹⁵. Extreme discharges entrain sediment through slope wash, rilling and gullying^{4,12}. In lower-severity burns, discontinuous runoff generation limits sediment yields from post-fire storm events¹⁶. (2) Loss of root strength a few years after tree mortality promotes shallow landslides¹⁷, usually during major winter storms. Storms that strike severe burns often produce thick, bouldery, charcoal-rich debris-flow deposits, but large debris flows may exist as only thin marginal facies at some alluvial fan sites^{4,12}. Discrete layers of charred forest litter provide another stratigraphic indicator of fire. These well-preserved burned soil surfaces imply rapid burial by post-fire sediments before bioturbation. We define large fire-related events ('large events') conservatively as debris-flow units with abundant coarse angular charcoal (Supplementary Fig. 1) that are coarser grained than other units in a stratigraphic section, and comprise at least 20% of its thickness; most overlie burned soil surfaces. Deposits that do not meet these criteria but that contain abundant charcoal and (or) overlie burned surfaces are considered small fire-related events ('small events'). Multiple events that occurred hours to several years after a single fire may appear as a single thick deposit that would be recognized as a 'large event', and still represent a major geomorphic response to a severe fire.

Individual charcoal fragments were AMS ¹⁴C-dated, using seeds, needles and twigs when available to reduce the probability of dating a sample much older than the fire in which it burned (yielding a large 'inbuilt age'). We obtained 133 dates from 33 stratigraphic sections (sites) in 32 different fans, with 1–15 dates per site (Supplementary Table 1). Calibrated probability distributions¹⁸ for 91 fire-related sedimentation events were summed to produce a probability spectrum (Fig. 2); multiple dates for the same fire-related event and stratigraphically inverted ages (Supplementary Table 2) were excluded. Small events dominate the overall record and are primarily responsible for major peaks except at ~750–1,000 and 1,700 calibrated calendar years before present (cal. yr BP). At-a-site frequency values were calculated to represent the frequency of fire-related events in a single fan's basin. Deposits of single events usually cover only part of a fan, and low- to moderate-severity fires often result in little or no deposits, thus at-a-site frequency of fire-related sedimentation is much lower than

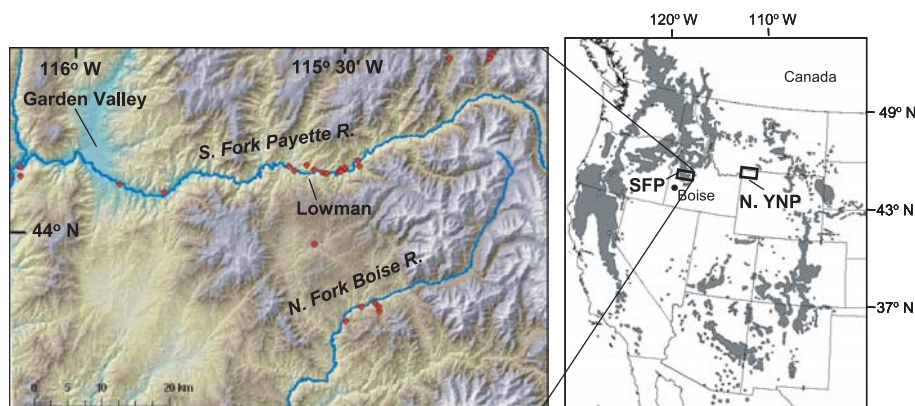


Figure 1 The SFP and northern Yellowstone National Park (YNP) field areas. Right, grey shading shows distribution of *Pinus ponderosa* (ponderosa pine) in the western USA and Canada (ponderosa pine distribution available at <http://climchange.cr.usgs.gov/data/>

atlas). Left, red dots on shaded relief map of the SFP study area show locations of alluvial fan stratigraphic sites. Elevations range from ~850 m in Garden Valley to >2,000 m on the eastern border of the map.

fire frequency in the associated basin. Nonetheless, at-a-site recurrence intervals for small sedimentation events are as low as 33–80 yr. Maxima in small events (Fig. 2a) at ~350, 1,200–1,500, 2,800–3,000, ~5,000 and 6,800–7,400 cal. yr BP correspond with peak at-a-site frequency over the study area, indicating periods of frequent low- to mixed-severity fires that produce limited sedimentation.

In contrast, about 24–27% of the total dated fan thickness was emplaced by only 9 large fire-induced debris-flow events ~1,000–800 cal. yr BP (Fig. 3). The importance of infrequent large events is supported by Holocene-average sediment flux estimates for Idaho batholith basins an order of magnitude greater than modern rates measured without major debris-flow events¹¹. Most fans in the lower-elevation ponderosa pine environment contain debris-flow units interbedded with multiple thin sheetflood deposits. Thick (0.5–2 m) fire-related debris-flow deposits dominate higher-elevation fans, consistent with a greater importance of severe stand-replacing fires in mixed-conifer forests.

We compare SFP fire-related sedimentation with a similar record from high-elevation forests in northern Yellowstone National Park

(Fig. 1), a cooler and wetter environment than the SFP area, where dense lodgepole-pine-dominated forests burn primarily in large, severe fires with recurrence intervals of 150–400 yr (refs 12, 14, 19). Notably, maxima in small-event probability (Fig. 2a) in the SFP area generally correspond with prominent minima in fire-related sedimentation in Yellowstone (~350–500, 1,200–1,400 and 2,800–3,000 cal. yr BP), suggesting a regional millennial-scale climatic control. In Yellowstone, these minima in fire-related sedimentation are associated with colder, effectively wetter, climates and increased runoff in streams, including during the Little Ice Age (LIA)¹². LIA climate was not uniformly cold in either time or space, but glacial advances occurred at ~650–100 cal. yr BP in the northern Rocky Mountains^{6–8}, with colder than present conditions in many Northern Hemisphere localities^{8–10}.

We infer that over much of the LIA, colder conditions maintained high canopy moisture, inhibiting stand-replacing fires in both Yellowstone lodgepole pine forests and SFP ponderosa pine forests^{12–14}. Concurrently, effectively wetter conditions increased understory grass growth and provided fuel for low-severity burns¹³ in the SFP, where summers are warmer and drier than in Yellowstone. In ponderosa pine forests, non-lethal surface fires typically occur when several cool, moist years produce abundant fine fuels, followed shortly by a mild to moderate drought year^{13,20–22}. Without prolonged severe droughts, wet–dry cycles on annual to multiyear timescales promote frequent fires in the SFP, and an increase in storm frequency during wetter intervals would also augment the number of small fire-related sedimentation events recorded in alluvial fans.

In contrast, intervals of stand-replacing fires and large debris-flow events are largely coincident in SFP ponderosa pine forests and Yellowstone, most notably during the ‘Medieval Climatic Anomaly’ (MCA), ~1,050–650 cal. yr BP (ref. 23; Fig. 2b). Although often called the ‘Medieval Warm Period’, times of anomalous Northern Hemisphere warmth are confined to shorter intervals within this period^{10,23–25}. In the western USA, the MCA included widespread, severe multidecadal droughts^{23,26}, with increased fire activity across diverse northwestern conifer forests^{12,13}. Some unusually

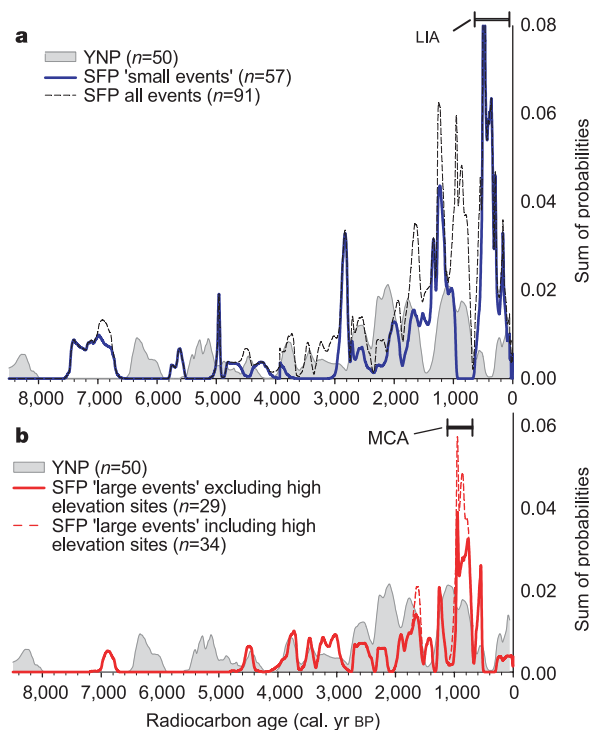


Figure 2 Summed calibrated probability distributions for radiocarbon ages on fire-related sedimentation events in the SFP Idaho area (this study) and in Yellowstone National Park¹² (YNP, grey-filled curves). Probability distributions were smoothed using a 100-yr running mean to reduce the influence of short-period variations in atmospheric ¹⁴C, but retain major probability peaks representing the most probable age ranges. Calibrated years before present (cal. yr BP) are equal to the number of calendar years before AD 1950. The general trend of decreasing probability over time, and overall lower probability values before ~4,000 cal. yr BP reflect lesser exposure and preservation of older deposits.

a, SFP ‘small events’ (blue line) in Idaho are thin deposits probably related to low- or moderate-severity burns; note correspondence of peaks with minima in YNP fire-related sedimentation, and major peak during the ‘Little Ice Age’ (LIA), ~650–50 cal. yr BP. (Fewer near-surface deposits in the 400 cal. yr BP–present range were selected for dating because of bioturbation and large uncertainties in ¹⁴C calibration.) **b**, SFP ‘large events’ are major debris-flow deposits probably related to severe fires; note correlation with the YNP record, and prominent peak in large-event probability during the ‘Medieval Climatic Anomaly’ (MCA), ~1,050–650 cal. yr BP.

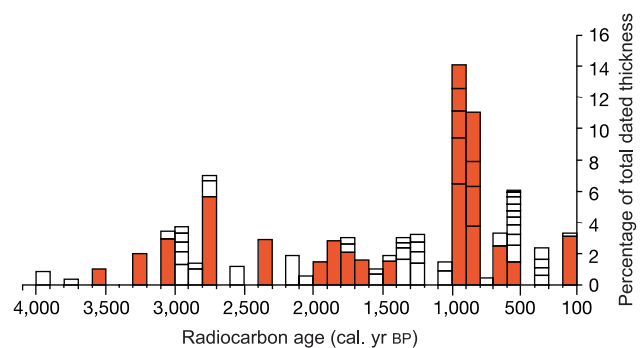


Figure 3 Variations in the thickness of fire-related deposits over time. Relative stratigraphic thickness of dated fire-related fan deposits in the SFP are expressed as a percentage of the total thickness for all dated deposits and plotted by calibrated age interval. Individual deposit thicknesses are shown by stacked bar segments and are calculated as the average of maximum and minimum possible thickness. This range in thickness results from measured variations in deposit thickness and from uncertainty in defining deposit boundaries. Average range of uncertainty in deposit thickness is ±10%. Red bar segments represent ‘large events’ based on criteria described in the text. Because of calibration problems for very young radiocarbon samples, deposits <100 cal. yr BP are not shown, and data older than 4,000 cal. yr BP are too sparse to be informative. Across the study area, ~50% of the total measured thickness of fan sediments deposited over the past 2,000 yr is probably or possibly related to fire.

wet decades also occurred, at least locally²⁷. Despite high fire frequency in the LIA, large SFP debris-flow events during the MCA account for most of the dated fire-related sediment produced over the last millennium (Fig. 3).

Although climate-induced changes in forest composition could drive changes in fire regimes, identification of carbonized macrofossils in fan deposits indicates that at least after ~4,500 cal. yr BP, the same conifer species comprising present forests existed in tributary basins (Supplementary Table 1; Supplementary Fig. 2). The inferred changes in fire frequency and severity may have stemmed in part from changes in relative species abundance—for example, infilling of open ponderosa pine stands by Douglas-fir as in recent decades in central Idaho³—but we cannot resolve such spatial changes in forest composition. At site LO1, now characterized by ponderosa pine and minor Douglas-fir, only Douglas-fir was identified at 7,400–6,800 cal. yr BP. Nonetheless, frequent small sedimentation events during this interval suggest a low- to mixed-severity fire regime, similar to ponderosa-dominated stands in the 1600s–1800s (ref. 3).

Along with previous studies covering shorter timescales, our results illustrate how superposed centennial- to millennial-scale climate variations influence fuel conditions, wildfire, and fire-related erosional events. Within multi-centennial cool-moist periods such as the LIA, moderate annual to multi-annual droughts produced frequent fires in xeric SFP ponderosa pine forests, controlled primarily by growth and desiccation of grass and other fine surface fuels. Concurrently, a cooler and effectively moister regional climate inhibited most fires in the high-elevation forests of Yellowstone.

In contrast, extreme droughts of up to multi-decadal length were associated with severe stand-replacing fires and large fire-related erosional events in both the SFP and Yellowstone during the ~400-yr MCA. Denser stands possibly aided these severe fires; perhaps prolonged droughts limited grass growth and surface fires, allowing survival of understory trees as ladder fuels^{1–3,13}. Alternatively, decades of unusually wet summers allowed densities to increase. In any case, extreme drought promotes critically low canopy moisture¹³ and is associated with MCA stand-replacing fires, despite the absence of fire suppression or land-use effects. Relatively severe droughts within the LIA may have resulted in some stand-replacing fires, as suggested by tree-ring studies in other ponderosa pine forests^{22,28,29}. Peak SFP fire-related sedimentation at ~350 cal. yr BP (~AD 1600) may partly reflect a prolonged drought in the late AD 1500s²⁶. Nonetheless, limited geomorphic response indicates that fires in the LIA were overall less severe than in either the MCA or recent decades.

Instrumental and proxy records reveal a rapid rise in Northern Hemisphere temperatures over the last century^{5,8,9} that is probably in part anthropogenic²⁴, and current warmth is probably greater than in the MCA²⁵. Warming of 0.6–2 °C has affected much of the western USA during the last century, and has been accompanied by a 5–20% decrease in precipitation³⁰ in northern ponderosa pine environments. Fire-season Palmer drought severity index correlates strongly with burn area in both central Idaho and Yellowstone^{12,20}, and shows a significant ($P = 0.01$) increase in drought severity between 1895 and 2002. In addition, twentieth-century fire suppression and to a lesser extent grazing have limited surface fires, allowing seedling survival and increased densities in SFP ponderosa pine stands³. Alteration of fuel conditions combined with increased incidence of extreme drought have probably produced the recent spate of extensive severe fires in the SFP and other northern ponderosa pine forests.

Fire management and ecological restoration strategies in ponderosa pine forests typically aim to prevent large stand-replacing fires by reproducing pre-settlement conditions with low tree densities^{1,3}. Climate exerts a powerful control on fire regimes, however, and the rapidity and magnitude of twentieth-century

global climate change is probably greater than has occurred for millennia^{24,25}. Efforts to return to fire regimes typical of a generally colder pre-settlement era will need to adapt to changing vegetation and fire activity in a warmer and drought-prone future. □

Received 17 February; accepted 15 September 2004; doi:10.1038/nature03058.

1. Covington, W. W. *et al.* Historical and anticipated changes in forest ecosystems of the inland west of the United States. *J. Sustain. For.* **2**, 13–63 (1994).
2. Agee, J. K. *Fire Ecology of Pacific Northwest Forests* (Island, Washington DC, 1993).
3. Steele, R., Arno, S. F. & Geier-Hayes, K. Wildfire patterns change in central Idaho's ponderosa pine-Douglas-fir forest. *West. J. Appl. For.* **1**, 16–18 (1986).
4. Meyer, G. A., Pierce, J. L., Wood, S. H. & Jull, A. J. T. Fires, storms, and erosional events in the Idaho batholith. *Hydrol. Process.* **15**, 3025–3038 (2001).
5. Mann, M. E., Bradley, R. S. & Hughes, M. K. Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. *Geophys. Res. Lett.* **26**, 759–762 (1999).
6. Carrara, P. E. Late Quaternary glacial and vegetative history of the Glacier National Park region, Montana. *US Geol. Surv. Bull.* **1902**, 1–64 (1989).
7. Luckman, B. H. The Little Ice Age in the Canadian Rockies. *Geomorphology* **32**, 357–394 (2000).
8. Grove, J. M. *The Little Ice Age* (Methuen, New York, 1988).
9. Pollack, H. N., Huang, S. & Shen, P. Climate change record in subsurface temperatures: a global perspective. *Science* **282**, 279–281 (1998).
10. Esper, J., Cook, E. R. & Schweingruber, F. H. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science* **295**, 2250–2253 (2002).
11. Kirchner, J. W. *et al.* Mountain erosion over 10 yr, 10 k.y., and 10 m.y. time scales. *Geology* **29**, 591–594 (2001).
12. Meyer, G. A., Wells, S. G. & Jull, A. J. T. Fire and alluvial chronology in Yellowstone National Park: Climatic and intrinsic controls on Holocene geomorphic processes. *Geol. Soc. Am. Bull.* **107**, 1211–1230 (1995).
13. Rollins, M. G., Morgan, P. & Swetnam, T. Landscape-scale controls over 20th century fire occurrence in two large Rocky Mountain (USA) wilderness areas. *Landscape Ecol.* **17**, 539–557 (2002).
14. Whitlock, C., Shafer, S. L. & Marlon, J. The role of climate and vegetation change in shaping past and future fire regimes in the northwestern US and the implications for ecosystem management. *For. Ecol. Manage.* **178**, 163–181 (2003).
15. Moody, J. A. & Martin, D. A. Comparison of soil infiltration rates in burned and unburned mountainous watersheds. *Hydrol. Process.* **15**, 2981–2993 (2001).
16. Lavee, H., Kutiel, P., Segev, M. & Benyamini, Y. Effect of surface roughness on runoff and erosion in a Mediterranean ecosystem: the role of fire. *Geomorphology* **11**, 227–234 (1995).
17. Schmidt, K. M. *et al.* The variability of root cohesion as an influence on shallow landslide susceptibility in the Oregon Coast Range. *Can. Geotech. J.* **38**, 995–1024 (2001).
18. Stuiver, M. & Reimer, P. J. Extended ¹⁴C data base and revised CALIB 3.0 ¹⁴C age calibration program. *Radiocarbon* **35**, 215–230 (1993).
19. Barrett, S. W. Fire regimes on andesitic mountain terrain in northeastern Yellowstone National Park, Wyoming. *Int. J. Wildl. Fire* **4**, 65–76 (1994).
20. Westerling, A. L., Gershunov, A., Brown, T. J., Cayan, D. R. & Dettinger, M. D. Climate and wildfire in the western United States. *Bull. Am. Meteorol. Soc.* **84**, 595–604 (2003).
21. Swetnam, T. W. & Betancourt, J. L. Mesoscale disturbance and ecological response to decadal climatic variability in the American Southwest. *J. Clim.* **11**, 3128–3147 (1998).
22. Veblen, T. T., Kitzberger, T. & Donnegan, J. Climatic and human influences on fire regimes in ponderosa pine forests in the Colorado Front Range. *Ecol. Appl.* **10**, 1178–1195 (2000).
23. Stine, S. In *Water, Environment and Society in Times of Climatic Change* (eds Issar, A. S. & Brown, N.) 43–67 (Kluwer, Dordrecht, 1998).
24. Crowley, T. J. Causes of climate change over the past 1000 years. *Science* **289**, 270–277 (2000).
25. Bradley, R. S., Hughes, M. K. & Diaz, H. F. Climate in medieval time. *Science* **302**, 404–405 (2003).
26. Woodhouse, C. A. & Overpeck, J. T. 2000 years of drought variability in the central United States. *Bull. Am. Meteorol. Soc.* **79**, 2693–2714 (1998).
27. Adams, K. Age and paleoclimatic significance of late Holocene lakes in the Carson Sink, NV, USA. *Quat. Res.* **60**, 294–305 (2003).
28. Brown, P. M., Kaufmann, M. R. & Shepherd, W. D. Long-term, landscape patterns of past fire events in a montane ponderosa pine forest of central Colorado. *Landscape Ecol.* **14**, 513–532 (1999).
29. Shinneman, D. J. & Baker, W. L. Nonequilibrium dynamics between catastrophic disturbances and old-growth forests in ponderosa pine landscapes of the Black Hills. *Conserv. Biol.* **11**, 1276–1288 (1997).
30. Karl, T. R., Knight, R. W., Easterling, D. R. & Quayle, R. G. Indices of climate change for the United States. *Bull. Am. Meteorol. Soc.* **77**, 279–286 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements Research funding was provided by the NSF, the University of New Mexico (UNM) and the UNM Department of Earth and Planetary Sciences. We thank L. Huckell for assistance with macrofossil identification; K. Pierce, B. Huckell and L. McFadden for comments on the manuscript; and S. Wood, T. Lite, L. Rockwell, S. Caldwell, C. North, K. Grover-Wier, K. Pierce and W. Andersen for discussions and field assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.P. (jpierce@unm.edu).

Contrasting origins of the upper mantle revealed by hafnium and lead isotopes from the Southeast Indian Ridge

Barry B. Hanan¹, Janne Blichert-Toft², Douglas G. Pyle³
& David M. Christie⁴

¹Department of Geological Sciences, San Diego State University, 5500 Campanile Drive, San Diego, California 92182-1020, USA

²Ecole Normale Supérieure de Lyon, Laboratoire des Sciences de la Terre, CNRS UMR 5570, 46 Allée d'Italie, 69364 Lyon Cedex 7, France

³School of Ocean and Earth Science Technology, University of Hawaii, 1680 East-West Road, POST 612A, Honolulu, Hawaii 96822, USA

⁴College of Oceanic and Atmospheric Sciences, Oregon State University, 104 COAS Admin Bldg, Corvallis, Oregon 97331-5503, USA

The origin of the isotopic signature of Indian mid-ocean ridge basalts has remained enigmatic, because the geochemical composition of these basalts is consistent either with pollution from recycled, ancient altered oceanic crust and sediments, or with ancient continental crust or lithosphere. The radiogenic isotopic signature may therefore be the result of contamination of the upper mantle by plumes containing recycled altered ancient oceanic crust and sediments¹, detachment and dispersal of continental material into the shallow mantle during rifting and breakup of Gondwana², or contamination of the upper mantle by ancient subduction processes^{3,4}. The identification of a process operating on a scale large enough to affect major portions of the Indian mid-ocean ridge basalt source region has been a long-standing problem. Here we present hafnium and lead isotope

data from across the Indian–Pacific mantle boundary at the Australian–Antarctic discordance region of the Southeast Indian Ridge, which demonstrate that the Pacific and Indian upper mantle basalt source domains were each affected by different mechanisms. We infer that the Indian upper-mantle isotope signature in this region is affected mainly by lower continental crust entrained during Gondwana rifting, whereas the isotope signature of the Pacific upper mantle is influenced predominantly by ocean floor subduction-related processes.

Geochemists have identified five types of mantle reservoirs on the basis of radiogenic isotope tracers⁵. One, the depleted asthenosphere mid-ocean ridge basalts (MORB) mantle source (DM⁶) is variably polluted on a regional scale by at least two of these isotopically distinct components^{7,8}. One, C-like⁸, is typified by intermediate ²⁰⁶Pb/²⁰⁴Pb, the other, EM-1-like and EM-2-like or DUPAL-like^{6–8}, can have very low ²⁰⁶Pb/²⁰⁴Pb. The marked contrast in the radiogenic isotope signatures between Indian and Pacific MORB is attributed to a greater proportion of an EM-1-like component, with very low ²⁰⁶Pb/²⁰⁴Pb and positive $\Delta^{207}\text{Pb}$ and $\Delta^{208}\text{Pb}$ (ref. 9), in the Indian upper mantle⁸. The origin of the EM-1-like material may represent ancient subcontinental lithospheric mantle (SCLM) or ancient recycled oceanic crust and sediment^{4,10–12}.

The known Indian Ocean plumes are too enriched in radiogenic Pb isotopes to account for the distinctly lower ²⁰⁶Pb/²⁰⁴Pb of Indian MORB relative to Pacific MORB⁴. The low trace-element abundances of SCLM, which require very large amounts for it to be an effective contaminant, and average ¹⁸⁷Os/¹⁸⁶Os ratios that are significantly lower than depleted MORB, also do not support SCLM as the origin of the EM-1 signature^{4,13}. The alternative remaining options for the Indian MORB EM-1 isotopic signature are thus sediment and altered oceanic crust accreted to the continent before being mixed into the upper-mantle MORB source², recycled material stored at a mantle boundary layer¹², or mantle wedge contamination by subducting slabs³. Here we present new Hf

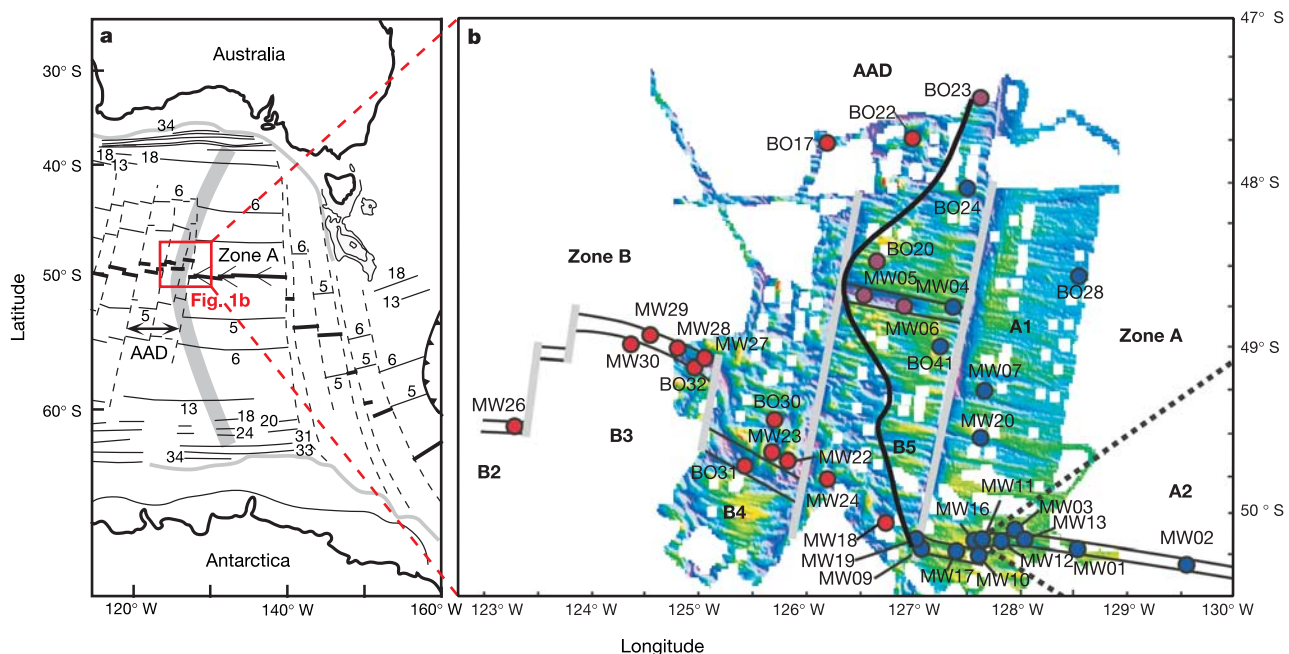


Figure 1 Regional setting of the AAD along the SEIR between Australia and Antarctica, schematic tectonic summary, and sample dredge locations. **a**, Transform fault zones (thin dashed lines) and magnetic lineations (solid lines) are shown. The broad grey curved line represents the regional depth anomaly¹⁶ that is at present associated with the AAD. The off-axis trace of the isotopic boundary lies within the eastern half of the depth anomaly (see Supplementary Information for the references used to compile Fig. 1a). **b**, Sample

locations for the sample data shown in Supplementary Table 1 are shown on the Seabeam bathymetry of the AAD region of the SEIR¹⁶: blue circles, Pacific-type; red circles, Indian-type; cross-hatched circles, transitional mantle sources. (Sample numbers with prefixes MW and BO correspond to the sample names in Supplementary Table 1) The ridge axis is bounded by the thin black lines. Fracture zones are represented by the thick grey lines. The approximate location of the AAD mantle boundary is indicated by the wavy black line.

and Pb isotope data for zero-age¹⁴ and off-axis basalt glasses from the Southeast Indian Ridge (SEIR)^{15,16} that support a model in which the shallow mantle in the Australian–Antarctic discordance (AAD) region was polluted by lower continental crust during rifting and breakup of Gondwana.

The AAD is the location of a geochemical and tectonic boundary between the Indian and Pacific upper-mantle provinces (Fig. 1). The deepest contiguous segments of the global mid-ocean ridge system lie within the AAD. The AAD and associated depth anomaly are an expression of cool upper mantle that has persisted since the separation of Australia from Antarctica during Gondwana breakup about 50 Myr ago. The origin of the depth anomaly has been variously attributed to a fragment of foundered SCLM, a stagnated subducted slab, or convective downwelling beneath the SEIR¹⁶.

The sharp mantle boundary occurs over a few tens of kilometres^{3,14–17}. Compared to zone A1–2 Pacific MORB from east of the AAD, zone B2–4 Indian MORB to the west have lower ²⁰⁶Pb/²⁰⁴Pb values combined with higher $\Delta^{208}\text{Pb}$ and $\Delta^{207}\text{Pb}$ values, and more radiogenic ⁸⁷Sr/⁸⁶Sr, all characteristics of strong influence from an EM-1 mantle source component. The range in ϵ_{Hf} is similar on both sides of the mantle discontinuity from about 123°–130° E (refs 3, 17), except for the zone B4 segment, which has the most ultra-depleted Hf isotope compositions reported so far for

MORB worldwide¹⁷. A step-like discontinuity is visible in the $\Delta\epsilon_{\text{Hf}}$ trend ($\Delta\epsilon_{\text{Hf}}$ correlates with increasing source depletion) from higher values (>1) on the Indian side of the boundary to lower values (<1) on the Pacific side (Supplementary Fig. 1). Basalts from zone B5, adjacent to the AAD boundary, derive from a mixture of Indian and Pacific mantle. These basalts have radiogenic isotope compositions transitional between zone B Indian and zone A Pacific MORB^{3,14,17}.

The C-like component, with intermediate Pb, Sr, Nd and Hf isotope ratios (for example, C⁸ or FOZO¹⁸ mantle components) relative to oceanic island basalts and MORB, is thought to derive from the deep mantle via hotspot mantle plumes⁵. The low-²⁰⁶Pb/²⁰⁴Pb EM-1-like component has isotope characteristics typical of recycled ancient oceanic sediment or SCLM^{11,19}. Binary-like arrays for the Pacific, Atlantic and Indian MORB subpopulations converge towards the intermediate component C at relatively higher ²⁰⁶Pb/²⁰⁴Pb, but are distinguished at low ²⁰⁶Pb/²⁰⁴Pb by differences in their relative proportions of the DM and EM-1 components⁸. In the AAD the contrast between Pacific and Indian MORB is striking in the ϵ_{Hf} –²⁰⁶Pb/²⁰⁴Pb diagram (Fig. 2a), in which there is no overlap between zone A and zone B MORB, except for the transitional, mixed-source basalts^{3,17}. Zone A basalts define a tight cluster with relatively low ϵ_{Hf} and ²⁰⁶Pb/²⁰⁴Pb and fall along a

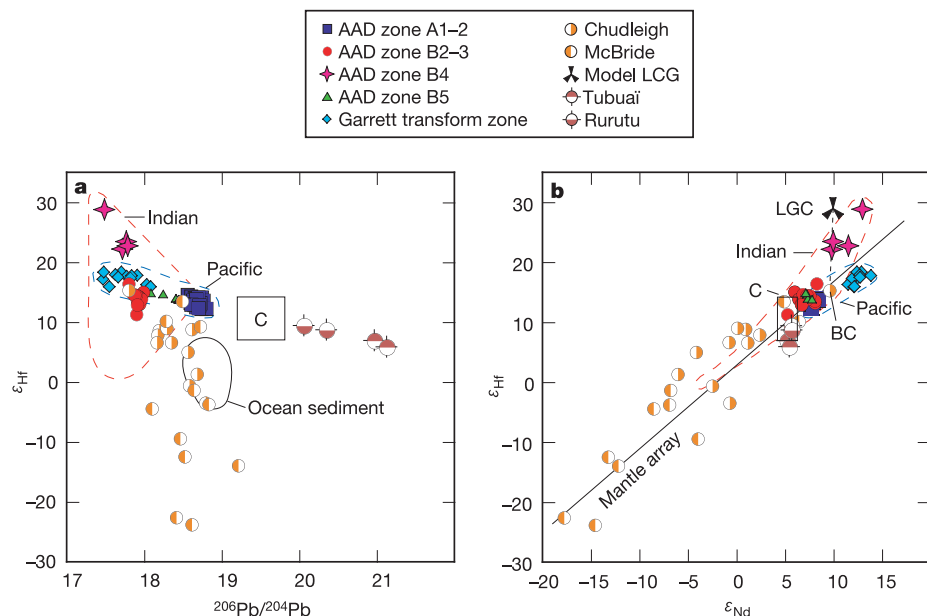


Figure 2 The ϵ_{Hf} , ϵ_{Nd} and ²⁰⁶Pb/²⁰⁴Pb variations of MORB from the AAD region of the SEIR. **a**, Pacific MORB from zone A1–2 fall on an array with ultra-depleted MORB²⁰ (J. Blichert-Toft and C. Hemond, personal communication) from the Garrett transform zone, the common component C⁸, and Pacific HIMU ocean island basalts²² from Tubuai and Rurutu. Indian zone B2–4 MORB define a separate array oblique to the Pacific array and include lower crustal granulites from Australia^{25,26}. The field for Pacific and Indian oceanic sediment^{27,28} overlaps with the Australian xenoliths. The fields for Pacific and Indian fields include the AAD zone B2–4 and zone A1–2 data and the MORB data from Chauvel and Blichert-Toft²¹. Note that in **b** the ultra-depleted Garrett transform zone and zone B4 basalts have completely overlapping ranges in ϵ_{Nd} . In a plot of ²⁰⁶Pb/²⁰⁴Pb versus ϵ_{Nd} , the Pacific and Indian trends do not cross, but intersect at the depleted ends of their arrays. This indicates that the two sources had similar time-integrated histories with respect to Sm/Nd, but marked differences in Lu/Hf fractionation. In this ²⁰⁶Pb/²⁰⁴Pb– ϵ_{Hf} projection the AAD zone A and other Pacific MORB²¹ define a linear trend but the AAD zone B2–4 and other Indian MORB²¹ together define a triangular shaped field that narrows and converges towards the C component⁸. In contrast, the Indian and Pacific trends are both linear-like and converge towards the C component in Pb isotope space. The range in ϵ_{Hf} increases with decreasing ²⁰⁶Pb/²⁰⁴Pb consistent with the idea that the low ²⁰⁶Pb/²⁰⁴Pb component in Indian MORB results from pollution of the regular upper-mantle MORB

source with a lower crustal granulite (LCG) component with a wide range of time-integrated Lu/Hf. The linear-like array shown by the AAD zone B2–4 data, within the Indian field, is dominated by the local heterogeneity. In Pb–Pb isotope correlations, the global Indian MORB population also shows relatively more dispersion (that is, heterogeneity) in ²⁰⁷Pb/²⁰⁴Pb and ²⁰⁸Pb/²⁰⁴Pb at a given low ²⁰⁶Pb/²⁰⁴Pb than the Pacific MORB population, consistent with a larger proportion of EM-1-like continental crustal component in the Indian upper-mantle MORB source. **b**, The Indian ultra-depleted MORB from zone B4 plot on or above the ϵ_{Nd} – ϵ_{Hf} mantle array line of Chauvel and Blichert-Toft²¹, whereas the Pacific Garrett transform zone ultra-depleted MORB fall on or below it. In general, except for the transitional zone-B5 mixed-source data, Pacific MORB have lower ϵ_{Hf} at the same ϵ_{Nd} than do Indian MORB. The black symbol represents the model composition of LCG depleted by early melting during the initiation of rifting 50 Myr ago. Sample BC, a mafic garnet granulite from the Chudleigh volcanic province, was used for the initial ϵ_{Hf} and ϵ_{Nd} isotope compositions^{25,26}. Values for Lu/Hf of 3.0 and Sm/Nd of 0.5 were assumed for the residue after melting. These parent/daughter trace-element ratios are in the range observed for the garnet granulite xenoliths from the Kilbourne Hole²⁹. The model data point (LCG), representing the residue, evolves about 14 ϵ_{Hf} units compared to less than one ϵ_{Nd} unit in 50 Myr. The dashed line connects the model LCG point and the starting material, sample BC.

typical Pacific MORB–ocean island basalt trend defined by the most depleted Pacific MORB from the Garrett transform zone²⁰ (J. Blichert-Toft and C. Hemond, personal communication), regular Pacific MORB²¹, the intermediate component C⁸ and Pacific HIMU⁶ (an enriched mantle component with high $\mu = {}^{238}\text{U}/{}^{204}\text{Pb}$) oceanic island basalts²². In contrast, zone B2–4 Indian MORB define a prominent negative trend oblique to and cross-cutting the Pacific trend. All reported Pacific MORB, including the ultra-depleted basalts from the Garrett transform zone, plot very near ($\Delta\epsilon_{\text{Hf}} < 1$) or below the mantle array in $\epsilon_{\text{Nd}}-\epsilon_{\text{Hf}}$ space, whereas the ultra-depleted basalts from zone B4 in the AAD and most Indian MORB plot on or above it ($\Delta\epsilon_{\text{Hf}} > 1$) (Figs 2b and 3).

The contrast between the Indian and Pacific isotope signatures could result from long-term differences in parent/daughter ratios related to the relative timing of melting events in an originally homogeneous upper mantle. Alternatively, the isotopic contrasts may be interpreted as evidence of melting of lithologically heterogeneous upper mantle. For example, the isotopic compositions of MORB and associated abyssal peridotites from the Southwest Indian Ridge show that a component other than peridotite is required to explain the enriched end of the range of variation in ϵ_{Nd} observed in these basalts²³. This component may be pyroxenite or eclogite²³ having a lower melting temperature than peridotite. The contrasting Indian and Pacific linear MORB trajectories in multi-isotope space are consistent with pollution and mixing of different recycled components within the upper-mantle MORB source regions over time⁸. In the ${}^{206}\text{Pb}/{}^{204}\text{Pb}-\epsilon_{\text{Hf}}$ diagram, zone A MORB fall along the Pacific trend. A two-component mantle consisting of pyroxenite or eclogite with a C-like isotope signature embedded in peridotite with a Garrett-transform-zone-like ultra-depleted isotope signature could yield first melts with zone-A-like composition, and, upon further melting, Garrett transform zone basalts.

Because the mineralogy of the Indian and Pacific MORB mantle sources are expected to be similar (pyroxene initially on the liquidus and similar trace-element abundance ratios), mantle mixing trends are linear rather than hyperbolic²⁴, implying that the enriched component in the Indian source should lie on the high- ${}^{206}\text{Pb}/{}^{204}\text{Pb}$ extension of the linear array defined by the zone B2–4 data. Because the zone B2–4 array cuts across the Pacific trend, the enriched component of the Indian trend must have an isotope signature

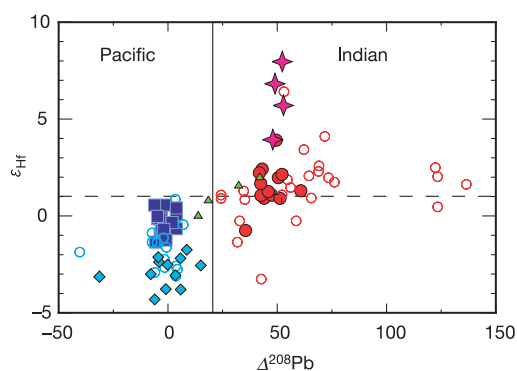


Figure 3 Contrasting $\Delta {}^{208}\text{Pb}$ and ϵ_{Hf} isotopic variations for MORB from the AAD (this work) and the Indian Ocean basin²¹ in general, versus the Pacific Ocean basin²¹. Pacific MORB have $\Delta {}^{208}\text{Pb} < 20$, while Indian MORB have $\Delta {}^{208}\text{Pb} > 20$. In terms of ϵ_{Hf} , the Pacific data are confined to the southwest quadrant of the diagram with $\epsilon_{\text{Hf}} < 1$. Indian MORB show a large range in ϵ_{Hf} , both positive and negative, but for the most part plot in the northeast quadrant with $\epsilon_{\text{Hf}} > 1$. The distinction between the Pacific and Indian MORB derives from pollution of the mantle source of the latter by continental crust. The light blue and red open circle symbols are Pacific and Indian MORB²¹, respectively. Other symbols are the same as in Fig. 2.

distinct from the enriched C-like or HIMU signature observed for the Pacific. Possible end-members include lower crust, such as represented by mafic granulite xenoliths from the nearby Chudleigh and McBride volcanic provinces of Australia^{25,26}, or oceanic sediment^{27,28}. The low- ${}^{206}\text{Pb}/{}^{204}\text{Pb}$ end of the Indian trend is also distinct from the Pacific trend. In contrast to the $\epsilon_{\text{Hf}}-{}^{206}\text{Pb}/{}^{204}\text{Pb}$ diagram, a plot of ${}^{206}\text{Pb}/{}^{204}\text{Pb}$ versus ϵ_{Nd} for the Pacific and Indian trends shows that these two arrays intersect at their depleted ends. Thus the two sources have had similar time-integrated histories with respect to Sm/Nd, but have experienced marked differences in Lu/Hf fractionation. Observations from a well-known incipient continental rift tectonic environment, the Rio Grande Rift, offer an explanation for the high- ϵ_{Hf} end-member and support the model for contamination of the Indian AAD upper mantle by lower crust.

Anomalous high $\Delta\epsilon_{\text{Hf}}$ values, such as those of the Indian MORB zone B4 samples, require a time-integrated Lu/Hf source composition substantially different from that of all other known oceanic basalt mantle sources, including both MORB and ocean island basalts. Garnet granulite xenoliths from the Kilbourne Hole, which represent residues from early magmatic processes during the initiation of the Rio Grande Rift around 25 Myr ago, have extremely high Lu concentrations and Lu/Hf ratios of 6.5–9.1 (ref. 29). The zone B4 ultradepleted component would require a comparatively much lower Lu/Hf ratio of only about 3 to develop its ϵ_{Hf} of about 30 from melting of Australian granulite-like material in 50 Myr (Fig. 2b). The high- ϵ_{Hf} low- ${}^{206}\text{Pb}/{}^{204}\text{Pb}$ end-member of the AAD zone B2–4 MORB array would thus be consistent with the melting of lower crustal granulite followed by the incorporation of the residues into the MORB mantle source during initial rifting about 50 Myr ago.

The origin of the low- ${}^{206}\text{Pb}/{}^{204}\text{Pb}$ EM-1 (and relatively higher- ${}^{206}\text{Pb}/{}^{204}\text{Pb}$ EM-2) signature in oceanic basalt may derive from continental crust or lithospheric mantle and/or sediment pollution. For the AAD, detachment and dispersal of continental lower crust into the shallow mantle during the rifting and breakup of Gondwana adequately accounts for the EM-1-like pollution of the zone B2–4 Indian MORB mantle source on the SEIR. We exclude recycled sediments as an alternative explanation for the EM-1 component in the Indian MORB mantle source for the following reason. Whereas strong EM-1 signatures are characteristic of Indian and South Atlantic mid-ocean ridges formed concomitantly with continental rifting and opening of these ocean basins, they are uncommon in the Pacific, which is an ocean basin not associated with continental rifting. Yet the Pacific is a region where recycling of oceanic crust and sediments occurs, largely in association with subduction and plume activity, and EM-2–C–HIMU isotopic signatures predominate. The contrast between the most depleted MORB from the Pacific Garrett transform zone MORB and Indian AAD zone B4 MORB requires a process that decouples the Lu–Hf and Sm–Nd systems in EM-1-like mantle material. This is a characteristic of rifting of lower crust, not of Pacific-style subduction and/or plume-related recycling. We also rule out pollution of the AAD Indian MORB source by subduction-modified mantle material forced into the Indian AAD upper mantle by a rising stagnated slab³. The ${}^3\text{He}/{}^4\text{He}$ of the ultra-depleted zone B4 basalts are among the highest reported for MORB away from hotspot influence (8.8 R/R_A ; D. Graham, personal communication), not low as would be expected for a subduction-related component³. Furthermore, the low ${}^{206}\text{Pb}/{}^{204}\text{Pb}$ ratios of the Indian AAD zone B2–4 MORB are not readily explained by pollution with components derived from Pacific altered oceanic crust and upper crustal sediment shed from Gondwana during the Cretaceous period, when the subducted oceanic crust was located just east of the Australian continental margin³⁰. The model for pollution of the AAD MORB source by lower crust may be applicable to other Indian MORB^{11,12}, to South Atlantic MORB³¹, the Arctic Gakkel ridge³², and to other

MORB that have formed in conjunction with continental rifting and where EM-1 component signatures occur. □

Received 8 March; accepted 17 September 2004; doi:10.1038/nature03026.

1. Storey, M. *et al.* Contamination of the Indian Ocean asthenosphere by the Kerguelen–Heard mantle plume. *Nature* **338**, 574–576 (1989).
2. Arndt, N. T. & Goldstein, S. L. An open boundary between lower continental crust and mantle: its role in crust formation and crustal recycling. *Tectonophysics* **161**, 210–212 (1989).
3. Kempton, P. D., Pierce, J. A., Barry, T. L., Langmuir, C. & Christie, D. M. Sr–Nd–Pb–Hf isotope results from ODP Leg 187: Evidence for mantle dynamics of the Australian–Antarctic discordance and origin of the Indian MORB source. *Geochem. Geophys. Geosyst.* **3**, doi:10.1029/2002GC000320 (2002).
4. Rehkämer, M. & Hofmann, A. W. Recycled ocean crust and sediment in Indian Ocean MORB. *Earth Planet. Sci. Lett.* **147**, 93–106 (1997).
5. Hofmann, A. W. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–229 (1997).
6. Zindler, A. & Hart, S. R. Chemical geodynamics. *Annu. Rev. Earth Planet. Sci.* **14**, 493–571 (1986).
7. Allègre, C. J., Hamelin, B., Provost, A. & Dupré, B. Topology in isotopic multislabs and the origin of mantle chemical heterogeneities. *Earth Planet. Sci. Lett.* **81**, 319–337 (1986/87).
8. Hanan, B. B. & Graham, D. W. Lead and helium isotope evidence from oceanic basalts for a common deep source of mantle plumes. *Science* **272**, 991–995 (1996).
9. Hart, S. R. A large-scale isotope anomaly in the Southern Hemisphere mantle. *Nature* **309**, 753–757 (1984).
10. Hamelin, B. & Allègre, C.-J. Large-scale regional units in the depleted upper mantle revealed by an isotope study of the South-West Indian Ridge. *Nature* **315**, 196–199 (1985).
11. Mahoney, J. J., Nicollet, C. & Dupuy, C. Madagascar basalts: tracking oceanic and continental sources. *Earth Planet. Sci. Lett.* **104**, 350–363 (1991).
12. Mahoney, J. J., White, W. M., Upton, B. G. J., Neal, C. R. & Scrutton, R. A. Beyond EM-1: lavas from Afanasy–Nikitin rise and the Crozet archipelago, Indian Ocean. *Geology* **24**, 615–618 (1996).
13. Pearson, D. G. *et al.* Archaean Re–Os age for Siberian eclogites and constraints on Archaean tectonics. *Nature* **374**, 711–713 (1995).
14. Pyle, D. G., Christie, D. M. & Mahoney, J. J. Resolving an isotope boundary within the Australian–Antarctic discordance. *Earth Planet. Sci. Lett.* **112**, 161–178 (1992).
15. Pyle, D. G., Christie, D. M., Mahoney, J. J. & Duncan, R. A. Geochemistry and geochronology of ancient southeast Indian and southwest Pacific seafloor. *J. Geophys. Res.* **100**, 22261–22282 (1995).
16. Christie, D. M., West, B. P., Pyle, D. G. & Hanan, B. B. Chaotic topography, mantle flow and mantle migration in the Australian–Antarctic discordance. *Nature* **394**, 637–644 (1998).
17. Hanan, B. B., Blichert-Toft, J., Christie, D. M. & Albarède, F. Ultra-depleted hafnium isotopes from Australian–Antarctic discordance MORB. *J. Conf. Abstr.* **5** (2), 478–479 (2000).
18. Hart, S. R., Hauri, E. H., Oschmann, L. A. & Whitehead, J. A. Mantle plumes and entrainment: isotopic evidence. *Science* **256**, 517–520 (1992).
19. Weaver, B. L. The origin of ocean island basalts end-member compositions: trace element and isotopic constraints. *Earth Planet. Sci. Lett.* **104**, 381–397 (1991).
20. Wendt, J. I., Regelous, M., Niu, Y., Hekinian, R. & Collerson, K. D. Geochemistry of lavas from the Garrett transform fault; insights into mantle heterogeneity beneath the eastern Pacific. *Earth Planet. Sci. Lett.* **173**, 271–284 (1999).
21. Chauvel, C. & Blichert-Toft, J. A hafnium isotope and trace element perspective on melting of the depleted mantle. *Earth Planet. Sci. Lett.* **190**, 137–151 (2001).
22. Salters, V. J. M. *et al.* Hf isotope constraints on mantle evolution Geochemical Earth Reference Model (GERM). *Chem. Geol.* **145**, 447–460 (1998).
23. Salters, V. J. M. & Dick, H. J. B. Mineralogy of the mid-ocean-ridge basalt source from neodymium isotopic composition of abyssal peridotites. *Nature* **418**, 68–72 (2002).
24. Hart, S. R., Gerlach, D. C. & White, W. M. A possible new Sr–Nd–Pb mantle array and consequences for mantle mixing. *Geochim. Cosmochim. Acta* **50**, 1551–1557 (1986).
25. Rudnick, R. L. & Goldstein, S. L. The Pb isotopic compositions of lower crustal xenoliths and evolution of lower crustal Pb. *Earth Planet. Sci. Lett.* **98**, 192–207 (1990).
26. Vervoort, J. D. *et al.* Hf–Nd isotopic evolution of the lower crust. *Earth Planet. Sci. Lett.* **181**, 115–129 (2000).
27. Ben Othman, D., White, W. M. & Patchett, P. J. The geochemistry of marine sediments, island arc magma genesis, and crust–mantle recycling. *Earth Planet. Sci. Lett.* **94**, 1–21 (1989).
28. Vervoort, J. D., Patchett, P. J., Blichert-Toft, J. & Albarède, F. Relationships between Lu–Hf and Sm–Nd isotopic systems in the global sedimentary system. *Earth Planet. Sci. Lett.* **168**, 79–99 (1999).
29. Scherer, E. E. *et al.* Lu–Hf geochemistry applied to dating Cenozoic events affecting lower crustal xenoliths from Kilbourne Hole, New Mexico. *Chem. Geol.* **142**, 63–78 (1997).
30. Gurnis, M., Mueller, R. D. & Moresi, L. Cretaceous vertical motion of Australia and the Australian–Antarctic discordance. *Science* **279**, 1499–1504 (1998).
31. Kamenetsky, V. S. *et al.* Remnants of Gondwanan continental lithosphere in oceanic upper mantle; evidence from the South Atlantic Ridge. *Geology* **29**, 243–246 (2001).
32. Soffer, G., Goldstein, S. L., Graham, D. W., Langmuir, C. H. & Michael, P. J. An arctic mantle domain boundary: Evidence from the Gakkel Ridge. *Geochim. Cosmochim. Acta* **68**, A692 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful for comments from D. Graham. We thank P. Têlouk for help with the Plasma 54. This work was supported by the National Science Foundation and the Institut National des Sciences de l'Univers.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.B.H. (Barry.Hanan@sdsu.edu).

The origin of the internal nostril of tetrapods

Min Zhu¹ & Per E. Ahlberg²

¹Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, P.O. Box 643, Beijing 100044, China

²Department of Evolutionary Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18A, SE-752 36 Uppsala, Sweden

The choana, a unique ‘internal nostril’ opening from the nasal sac into the roof of the mouth, is a key part of the tetrapod (land vertebrate) respiratory system. It was the first component of the tetrapod body plan to evolve, well before the origin of limbs, and is therefore crucial to our understanding of the beginning of the fish–tetrapod transition. However, there is no consensus on the origin of the choana despite decades of heated debate^{1–9}; some have claimed that it represents a palatally displaced external nostril^{4,6}, but others have argued that this is implausible because it implies breaking and rejoining the maxillary–premaxillary dental arcade and the maxillary branch of nerve V^{2,6}. The fossil record has not resolved the dispute, because the choana is fully developed in known tetrapod stem-group members^{8,10,11}. Here we present new material of *Kenichthys*, a 395-million-year-old fossil fish from China^{12–14}, that provides direct evidence for the origin of the choana and establishes its homology: it is indeed a displaced posterior external nostril that, during a brief transitional stage illustrated by *Kenichthys*, separated the maxilla from the premaxilla.

Most jawed fishes have anterior and posterior external nostrils but no connection between the nasal sac and mouth. The only living exception is the lungfishes, in which the posterior nostril is palatal, but this differs structurally from the tetrapod choana and is regarded as convergent^{2,3,6}. Among those sarcopterygian fishes (lobe-fins) with a complete maxillary–premaxillary arcade, the nasal region comes in two patterns. In fishes belonging to the Tetrapodomorpha, the tetrapod total group^{10–12}, there is a single external nostril that lies between lateral rostral and tectal (Fig. 1f, g). On the palate, the anterior end of the maxilla splits into two diverging processes that surround a choana^{3,15}. In non-tetrapodomorphs such as *Youngolepis*^{16,17}, which lack a choana (Fig. 1b, c), the anterior external nostril lies between the lateral rostral and tectal bones and the posterior nostril between the lateral rostral and lacrimal^{3,9}. On the palate, the maxilla and premaxilla meet in a simple butt joint. *Kenichthys campbelli*, from the Emsian Chuan-dong Formation of Yunnan, China, presents the only known intermediate between these two patterns.

Kenichthys is the earliest known and most phylogenetically basal tetrapodomorph^{11–14}, a conclusion that we confirm here on the basis of a new expanded character suite (Supplementary Information). The snout and cheek bones of *Kenichthys*, including 19 ethmosphenoid and 10 maxilla specimens, are not preserved in articulation but are very well preserved and show little individual variation in morphology. This allows their positional relationships to be reconstructed with reasonable certainty, even though it is not possible to reassemble individual skulls (Figs 2 and 3). The anterior nostril occupies its usual position between lateral rostral and tectal (Fig. 1a, d). Unusually, however, the ventral margin of the lateral rostral extends some way beyond the posterior end of the premaxilla (Figs 1a, d, e and 2e). The end of the premaxilla is narrow and capped by cosmine (dentine with a thin enamel covering), without a sutural surface for the maxilla, and the free ventral edge of the lateral rostral is also cosmine-covered rather than sutural (Fig. 2a–h, j). It seems that these margins were free and wrapped in epithelium rather than sutured to the maxilla as in other lobe-fins. The small

space that is bounded anteriorly by the posterior end of the premaxilla, and laterally by the ventral margin of the lateral rostral, lies immediately below the fenestra endochoanalis (Figs 2a, c and 3c, d). The posterior margin of the lateral rostral, which would carry the posterior nostril in non-tetrapodomorph lobe-fins, is entirely sutural and matches perfectly the anterior margin of the lacrimal (Fig. 2i). The final piece of the puzzle is provided by the maxilla (Figs 2k, l and 3f), which lacks the two diverging processes that frame the choana in other tetrapodomorph fishes, ending instead in a simple thin edge (Fig. 1a) and carrying no obvious sutural surface for the premaxilla.

Kenichthys clearly does not display either the primitive sarcopterygian condition with two external nostrils (Fig. 1b) or the tetrapodomorph condition with a single external nostril and a large choana framed by diverging processes of the maxilla (Figs 1g and 3g). We reconstruct the posterior nostril as opening in the upper lip, separating the maxilla from the premaxilla (Fig. 1d, e). No alternative reconstruction makes sense of all the morphological characteristics of the nasal region.

It is evident that there was no posterior external nostril on the suture between the lacrimal and lateral rostral, and no notch for a choana in the anterior end of the maxilla. If the maxilla is pushed forwards to contact the premaxilla and lateral rostral, we are left with no space for a posterior nostril or choana at all, except conceivably in a very mesial position internal to the maxillary-premaxillary arcade, which seems implausible. Such a reconstruction also fails to account for the clearly non-sutural character of the posterior end of the premaxilla and the free ventral margin of the lateral rostral.

The morphology of the maxilla itself also conflicts with this reconstruction. In ventral view (Fig. 2a, c), the free ventral margin of the lateral rostral runs at about 45° to the long axis of the head; the maxilla, in contrast, must have had an alignment closer to 30° to the long axis (because any other alignment results in an implausible head shape), so to suture with the ventral margin of the lateral rostral as in other tetrapodomorphs, the anterior end of the maxilla would have needed a distinct mesial deflection. In fact there is no

such deflection (Figs 1e and 2k, l). Furthermore, the dorsal margin of the maxilla shows a single continuous contact area anteriorly (Figs 2k, l and 3f), suggesting that this margin sutured with only one bone. Where a single long bone sutures with two shorter elements, such as the anterior part of the lacrimal suturing with the tectal and prefrontal, the boundary between the two sutures is usually marked by a slight step (Fig. 2i, red arrow). The overlapped area in the anterior part of the maxilla also excludes the possibility that the maxilla formed the ventral margin of the posterior nostril as in non-tetrapodomorph fishes.

We conclude that the anterior part of the maxilla sutured only

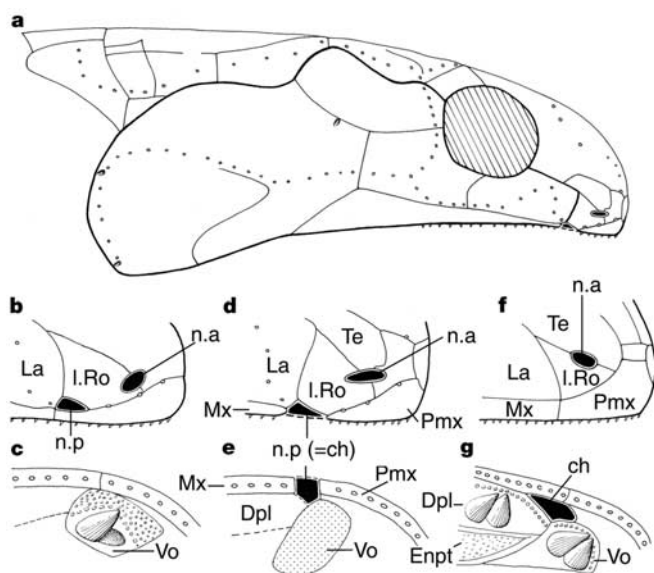


Figure 1 Nostril positions on the heads of sarcopterygian fishes. **a**, reconstruction of *Kenichthys campbelli* head and cheek in lateral view. **b, c**, *Youngolepis*. **d, e**, *Kenichthys*. **f, g**, *Eusthenopteron*. **b, d, f** are lateral views; **c, e, g** are ventral views. In **e**, the (unknown) vomer is represented by its attachment area on the ethmoid. Not to scale. ch, choana; Dpl, dermopalatine; Enpt, entopterygoid; La, lacrimal; I.Ro, lateral rostral; Mx, maxilla; n.a, anterior nostril; n.p, posterior nostril; Pmx, premaxilla; Te, tectal; Vo, vomere.

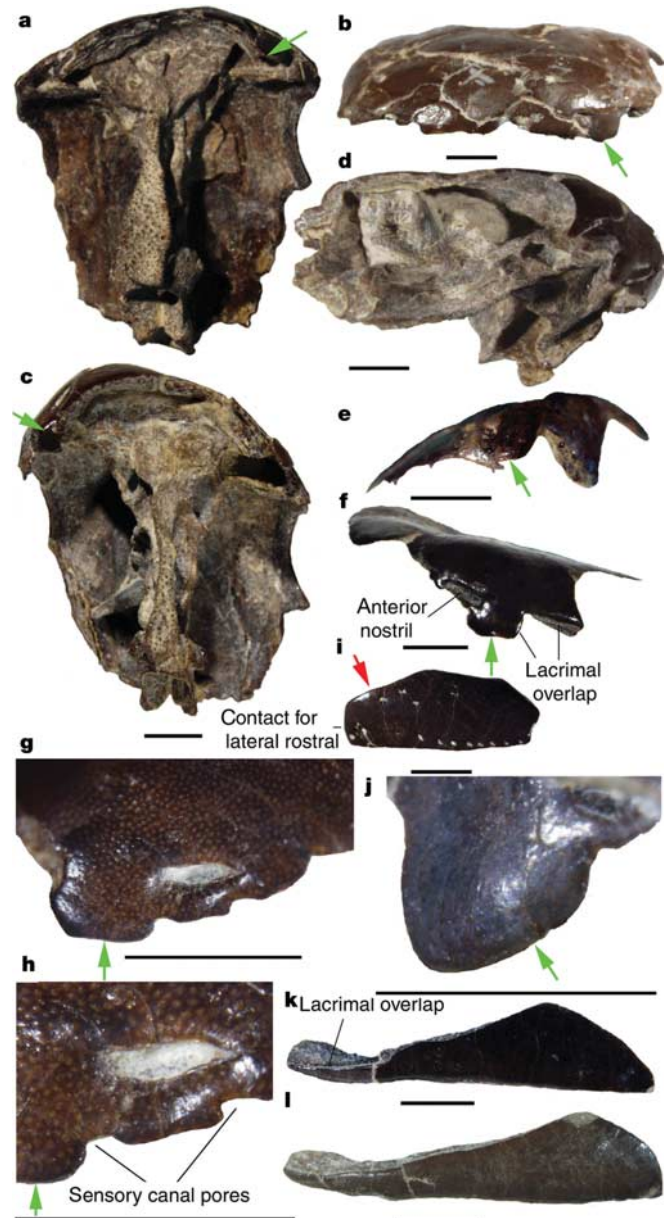


Figure 2 Photographs of *Kenichthys campbelli* specimens. **a, b**, V10493.61, ethmosphenoid in ventral (**a**) and anterior (**b**) views. **c, d**, V10493.60, ethmosphenoid in ventral (**c**) and lateral (**d**) views. **e, f**, V10493.58, ethmosphenoid in antero-lateral view. **g, h**, V10493.1, holotype, ethmosphenoid in antero-lateral view. **i, j**, V10493.81, lacrimal in external view. **k, l**, V10493.102, ethmosphenoid in anterior view. **m, n**, V10493.77, maxilla in external view. **o, p**, V10493.76, maxilla in external view. The green arrow indicates the free ventral margin of the lateral rostral; the red arrow in **i** indicates the slight 'step' marking the boundary of the tectal and prefrontal sutures. Scale bar, 2 mm.

with the lacrimal, and that the two bones ended anteriorly at the same level (Fig. 1a, d). The maxilla was separated from the premaxilla by a gap of about 1 mm, which communicated directly with the nasal capsule dorsally and was closed dorsolaterally by the free ventral edge of the lacrimal. This gap housed the posterior nostril of *Kenichthys* (Fig. 1a, d). The mesial extent of the nostril onto the palate is uncertain, because the vomer is unknown (apart from its attachment area) and the only known dermopalatine is incomplete and displaced (Fig. 3e), but it seems likely that it was relatively small and that the overall size of the two nostrils was similar.

The nostril morphology of *Kenichthys* forms a perfect intermediate between the non-choanate and choanate conditions. Given the phylogenetic position of *Kenichthys* as the most basal tetrapodomorph (which it maintains even without use of nostril characters¹¹), we infer that the posterior nostril was displaced ventrally at the internode below *Kenichthys*, causing the maxilla and premaxilla to become separated. At the internode between *Kenichthys* and rhizodonts (which have a single external nostril¹⁸, contrary to previous claims for two¹⁹), the maxilla and premaxilla re-established contact lateral to the posterior nostril, which by definition became a choana. *Kenichthys* thus demonstrates beyond reasonable doubt that the tetrapod choana is homologous with the posterior external nostril of fishes. This is further supported by recent work²⁰ disproving the supposed co-occurrence of a posterior external nostril and choana in porolepiforms³, previously one of the main stumbling blocks for that hypothesis^{1,3,4,6}. Furthermore, although *Kenichthys* shows only a single intermediate step, it strongly indicates that the transformation from external nostril to choana was essentially gradualistic in nature; there is nothing to suggest the need for a saltational leap

either before or after the *Kenichthys* node. This contrasts with the apparently abrupt changes in skull and limb structure at the *Panderichthys*–*Acanthostega* internode²¹, but cautions that the discovery of intermediates could similarly overturn our perception of those transformations.

Until the discovery of *Diabolepis*²², a major subject of dispute was the relationship between the tetrapod choana and the palatal posterior nostril of lungfishes^{2,4}. *Diabolepis*, the immediate sister group of the lungfishes within the Dipnomorpha²³, has a posterior nostril that is lateral to the premaxilla and therefore shows that the palatal position of the nostril in lungfishes and tetrapods is not homologous. However, as *Diabolepis* and the less crownward dipnomorph *Youngolepis* both have posterior nostrils that are very close to the lip, it might be that a ventral displacement of the posterior nostril is a shared derived character of Dipnomorpha + Tetrapodomorpha, with a further shift onto the palate accomplished independently through breaching of the maxillary–premaxillary arcade (Tetrapodomorpha) or loss of this arcade and inrolling of the lip (Dipnomorpha)⁶.

It is striking that the development of the nasal region in tetrapods seems in some ways to recapitulate the evolutionary transformation illustrated by *Kenichthys*. In fishes with two external nostrils, such as the paddlefish *Polyodon*²⁴, the premaxilla simply develops on the snout anteriorly to the maxillary prominence of the mandibular arch, without any involvement with the nasal sac. However, in human development, the nasal sac is initially open to the upper jaw margin, and this opening is only closed when the ‘central prominence’ (which will form the premaxilla²⁵) and the maxillary prominence grow together and fuse. Failure of this process results in a cleft lip, mirroring the gap between the premaxilla and maxilla in *Kenichthys*. This joining of the central and maxillary prominences resembles a recapitulation of the re-establishment of maxillary–premaxillary contact at the *Kenichthys*–rhizodont internode. The occurrence of cleft lip and palate are known to be associated with failure of *Bmp* and/or *Shh* + *ptc* expression^{25–27}, regulating the growth of the central and maxillary prominences as well as the epithelial breakdown that allows them to fuse, which indicates that components of this pathway might have been involved in regulating the evolution of the choana. In any event, it seems that a common defect in facial development refers back to an evolutionary event that occurred nearly 400 million years ago. □

Received 13 February; accepted 15 July 2004; doi:10.1038/nature02843.

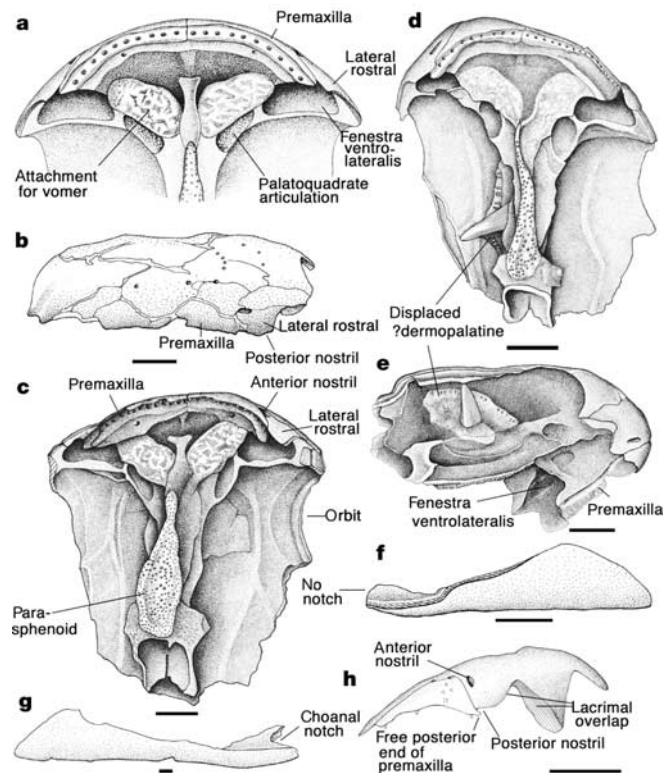


Figure 3 Morphology of *Kenichthys*. **a–f, h**, drawings of *Kenichthys campbelli* specimens. **a**, Restoration of ethmosphenoid in ventral view, based mainly on **c** and **d**. **b, c**, V10493.61, ethmosphenoid in anterior (**b**) and ventral (**c**) views. **d, e**, V10493.60, ethmosphenoid in ventral (**d**) and lateral (**e**) views. **f**, V10493.77, maxilla in external view. **g**, Maxilla of *Medoevia* (ref. 15), showing the diverging processes framing the choana. **h**, V10493.58, ethmosphenoid in antero-lateral view. Scale bar, 2 mm.

- Jarvik, E. On the structure of the snout of crossopterygians and lower gnathostomes in general. *Zool. Bidrag Uppsala* **21**, 235–675 (1942).
- Panchen, A. L. The nostrils of choanate fishes and early tetrapods. *Biol. Rev.* **42**, 374–420 (1967).
- Jarvik, E. *Basic Structure and Evolution of Vertebrates* Vol. 1 (Academic, London, 1980).
- Rosen, D. E. et al. Lungfishes, tetrapods, paleontology, and plesiomorphy. *Bull. Am. Mus. Nat. Hist.* **167**, 157–276 (1981).
- Maisey, J. G. Heads and tails: a chordate phylogeny. *Cladistics* **2**, 201–256 (1986).
- Panchen, A. L. & Smithson, T. R. Character diagnosis, fossils and the origin of tetrapods. *Biol. Rev.* **62**, 341–438 (1987).
- Forey, P. L. in *The Biology and Evolution of Lungfishes* (eds Bemis, W. E., Burggren, W. W. & Kemp, N.) 39–74 (Alan R. Liss, New York, 1987).
- Schultze, H.-P. in *Origins of the Higher Groups of Tetrapods: Controversy and Consensus* (eds Schultze, H.-P. & Trueb, L.) 29–67 (Cornell Univ. Press, Ithaca, 1991).
- Ahlberg, P. E. A re-examination of sarcopterygian interrelationships, with special reference to the Porolepiformes. *Zool. J. Linn. Soc.* **103**, 241–287 (1991).
- Long, J. A., Campbell, K. S. W. & Barwick, R. E. Osteology and functional morphology of the osteolepiform fish *Gogonasus andrewsae* Long, 1985, from the Upper Devonian Gogo Formation, Western Australia. *Rec. W. Aust. Mus. Suppl.* **53**, 1–89 (1997).
- Ahlberg, P. E. & Johanson, Z. Osteolepiforms and the ancestry of tetrapods. *Nature* **395**, 792–794 (1998).
- Chang, M. M. & Zhu, M. A new Middle Devonian osteolepidid from Qujing, Yunnan. *Mem. Assoc. Australas. Palaeontol.* **15**, 183–198 (1993).
- Zhu, M. & Wang, J.-Q. On the Early–Middle Devonian boundary in Qujing, Yunnan. *Acta Stratigr. Sin.* **20**, 58–63 (1996).
- Chang, M. M. & Yu, X. B. Reexamination of the relationship of Middle Devonian osteolepidids—fossil characters and their interpretations. *Am. Mus. Novitat.* **3189**, 1–20 (1997).
- Lebedev, O. A. Morphology of a new osteolepidid fish from Russia. *Bull. Mus. Natl. Hist. Nat. Paris* **17**, 287–341 (1995).
- Chang, M. M. *The Braincase of Youngolepis, a Lower Devonian Crossopterygian from Yunnan, South-western China* Thesis, Stockholm Univ. (1982).

17. Tong-Dzuy, T. & Janvier, P. Les Vertébrés du Dévonien inférieur du Bac Bo oriental (provinces de Bac Thai et Lang Son, Viêt Nam). *Bull. Mus. Natl. Hist. Nat. Paris* **12**, 143–223 (1990).
18. Johanson, Z. & Ahlberg, P. E. A complete primitive rhizodont from Australia. *Nature* **394**, 569–572 (1998).
19. Long, J. A. A new rhizodontiform fish from the Early Carboniferous of Victoria, Australia, with remarks on the phylogenetic position of the group. *J. Vert. Paleontol.* **9**, 1–17 (1989).
20. Clement, G. Evidence for lack of choanae in the Porolepiformes. *J. Vert. Paleontol.* **21**, 795–802 (2001).
21. Ahlberg, P. E., Clack, J. A. & Lukševičs, E. Rapid braincase evolution between *Panderichthys* and the earliest tetrapods. *Nature* **381**, 61–64 (1996).
22. Chang, M. M. & Yu, X. B. Structure and phylogenetic significance of *Diablichthys speratus* gen. et sp. nov., a new dipnoan-like form from the early Devonian of southeastern Yunnan, China. *Proc. Linn. Soc. N.S.W.* **107**, 171–184 (1984).
23. Zhu, M., Yu, X. B. & Ahlberg, P. E. A primitive sarcopterygian fish with an eyestalk. *Nature* **410**, 81–84 (2001).
24. Bemis, W. E. & Grande, L. Early development of the actinopterygian head. I. External development and staging of the paddlefish *Polyodon spathula*. *J. Morphol.* **213**, 47–83 (1992).
25. Ashique, A. M., Fu, K. & Richman, J. M. Endogenous bone morphogenetic proteins regulate outgrowth and epithelial survival during avian lip fusion. *Development* **129**, 4647–4660 (2002).
26. Hu, D. & Helms, J. The role of *sonic hedgehog* in normal and abnormal craniofacial morphogenesis. *Development* **126**, 4873–4884 (1999).
27. Zhang, Z. et al. Rescue of cleft palate in *Mx1*-deficient mice by transgenic *Bmp4* reveals a network of BMP and Shh signaling in the regulation of mammalian palatogenesis. *Development* **129**, 4135–4146 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. M. Chang for advice and discussions, M. Yang for artwork, and X. Lu for specimen preparation. This work was supported by the Special Funds for Major State Basic Research Projects of China and the Chinese Foundation of Natural Sciences. P.E.A. thanks the Royal Society and Chinese Academy of Sciences for supporting his visit to Beijing in 2002 through their exchange programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.Z. (zhumin@ivpp.ac.cn).

Ecosystem remodelling among vertebrates at the Permian–Triassic boundary in Russia

M. J. Benton¹, V. P. Tverdokhlebov² & M. V. Surkov²

¹Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK

²Geological Institute of Saratov State University, Ulitsa Moskovskaya 161, Saratov 410075, Russia

The mass extinction at the Permian–Triassic boundary, 251 million years (Myr) ago, is accepted as the most profound loss of life on record^{1–3}. Global data compilations indicate a loss of 50% of families or more, both in the sea^{1,2,4} and on land^{2,5}, and these figures scale to a loss of 80–96% of species, based on rarefaction analyses^{6,7}. This level of loss is confirmed by local and regional-scale studies of marine sections^{3,8}, but the terrestrial record has been harder to analyse in such close detail. Here we document the nature of the event in Russia in a comprehensive survey of 675 specimens of amphibians and reptiles from 289 localities spanning 13 successive geological time zones in the South Urals basin. These changes in diversity and turnover cannot be explained simply by sampling effects. There was a profound loss of genera and families, and simplification of ecosystems, with the loss of small fish-eaters and insect-eaters, medium and large herbivores and large carnivores. Faunal dynamics also changed, from high rates of turnover through the Late Permian period to greater stability at low diversity through the Early Triassic period. Even after 15 Myr of ecosystem rebuilding, some guilds were apparently still absent—small fish-eaters, small insect-eaters, large herbivores and top carnivores.

At a time when there is so much focus on global change and threats to biodiversity, it is surprising how little was known about the Permian–Triassic boundary (PTB) event in 1990 (refs 1, 2, 9). Over the past 15 years, our understanding of this mass extinction has become focused in terms of the timescale (perhaps lasting for 500,000 years (refs 3, 10)), the cause (probably associated with massive outpourings of basalt lava, the Siberian Traps, triggering global warming and anoxia, and possibly a runaway greenhouse effect associated with repeated release of gas hydrates^{2,8–12}), and the nature of the event and the immediate recovery phase (mass extinction followed by rapid turnover of weedy species during the phase of maximum anoxia, and then slow rebuilding of ecosystems^{3,8,9}).

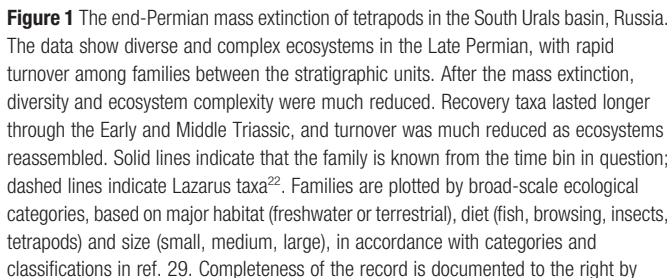
The Permian–Triassic succession of the South Urals is about 6 km thick, thinning to 1–2 km in the Moscow basin^{13–15}, and it is subdivided into 13 successive stratigraphic units (Fig. 1). These units are recognized in the field by changes in sedimentary rock type (svitas), and by particular fossil assemblages (gorizonts); they are correlated with each other, and with the global standard, by means of palynomorphs and ostracods^{13–15}. The age range covers the Kazanian and Tatarian stages of the Late Permian and the Induan to Ladinian stages of the Early and Middle Triassic, a total time span of 25–30 Myr.

The Late Permian to Triassic succession in the South Urals starts with a marine episode in the Kazanian, represented by 200 m of limestone, mudstone and halite, followed by about 1 km of river-deposited mudstone and sandstone. The continental succession extends with relatively continuous deposition from the late Kazanian to the Ladinian (Middle Triassic), and consists of repetitions of four main facies types: mudflats, sandy distributary channels, small gravelly channels and large gravelly channel fluvial systems¹³. The basalmost Triassic is marked by thick sandstone units that document a marked, but short-lived, change in sedimentation style to large gravelly channels, with boulders of more ancient rocks, up to 1 m across, swept down from the Ural mountains. These thick conglomerate units were deposited in large-scale alluvial fans that were part of a much larger terminal fan, about 350–400 km in width.

The abrupt change in the size of the basin and the incoming of coarse-grained alluvial fans all along the western margin of the Urals probably resulted from a peak in mountain-building activity in the core of the Urals, and a change at the PTB towards a more arid climate, with higher sediment yield and greater peak discharges in a drainage basin with reduced vegetation cover¹³. These massive changes in style of sedimentation at the PTB have been seen independently in the continental Karoo succession in South Africa¹⁶ and Australia¹⁷. The changes have been linked to the Siberian basalt eruptions and the consequent marked global warming and acid rain. The acid rain may have killed off the vegetation on land, and soils were stripped from the landscape and swept down rivers on to the plains, and eventually into the sea^{2,8,9}. Mountain uplift and soil stripping, rather than increased rainfall, lies behind the switch from low-energy rivers and cyclical deposition in the latest Permian to massive erosion at the base of the Triassic. Environments and sedimentation styles reverted to pre-PTB conditions higher in the Lower Triassic succession.

The range chart of tetrapods in the Late Permian and Triassic of the South Urals (Fig. 1) shows diverse ecosystems in the Late Permian¹⁵. In the rivers and lakes, four to seven genera of small, medium and large aquatic tetrapods ('amphibians') fed on the abundant thick-scaled bony fishes and rarer freshwater sharks and lungfishes. On the wooded banks were 5–11 genera of terrestrial tetrapods ('reptiles'), ranging in size from tiny insect-eaters to rhino-sized plant-eating pareiasaurs and the wolf-sized to bear-sized sabre-toothed gorgonopsians that fed on them. During the 17–18 Myr of the Kazanian and Tatarian, there was considerable turnover of genera and families through the six time zones (Fig. 2).

The percentage extinction of families at the end of the Permian



By the end of the sampling period, ecosystems were again

Analysis of the data shows that the patterns cannot be explained simply as artefacts of sampling. Plots of numbers of genera and families against numbers of localities and specimens (Fig. 3a, b) show no correlation: if anything, time bins with large numbers of localities and specimens are associated with low-diversity faunas and vice versa. Further, when the distributions of generic and

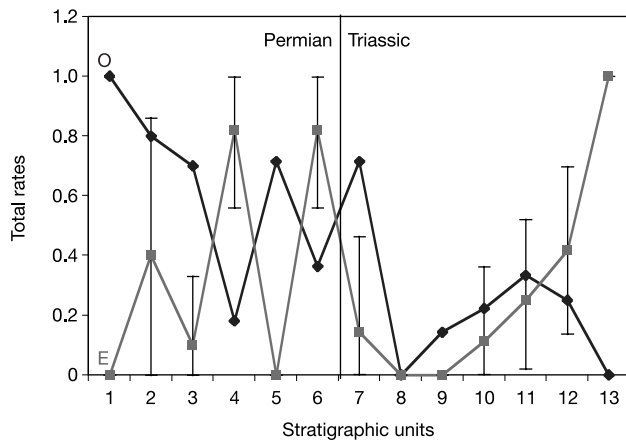


Figure 2 Turnover of tetrapod families through the Late Permian and Early Triassic in the South Urals basin, Russia. Rates are high and variable through the Late Permian, but they are depressed after the end-Permian mass extinction, and pick up slowly thereafter during the Early Triassic recovery phase. Rates of origination (O, diamonds) and extinction (E, squares) are percentage metrics based on all taxa (including Lazarus taxa, but excluding singleton families—families known from a single species or single locality) known from a time bin. Stratigraphic units are the successive svitas of the Upper Permian (1, Osinovskaya; 2, Belebey; 3, Bolshekinelskaya; 4, Amanakskaya; 5, Malokinelskaya/Vyakovskaya; 6, Kutulukskaya/Kulchomovskaya), Lower Triassic (7, Kopanskaya; 8, Staritskaya; 9, Kzylsaiskaya; 10, Gostevskaya; 11, Petropavlovskaya) and Middle Triassic (12, Donguz; 13, Bukobay). Binomial 95% confidence intervals³⁰ are shown for the percentage extinction metrics (confidence intervals are of similar magnitude for the percentage origination metrics, but are omitted for clarity).

familial diversity through time are compared with the distributions of numbers of sites and numbers of specimens in each time bin (Fig. 3c), there is no apparent tracking. Peaks and troughs in the diversity data do not match peaks and troughs in richness of the fossil record. And, crucially, the time of diversity decline across the PTB corresponds to a rising trend in numbers of sites and specimens.

Three sampling standardization protocols were also applied (see Methods) to assess whether the patterns of apparent diversity, and extinction and origination rates, could be determined by sample size. Five of the stratigraphic units are represented by small ($n < 50$) sample sizes, namely the Osinovskaya, Belebey, Bolshekinelskaya, Gostevskaya and Bukobay svitas, of which only the Gostevskaya falls near the PTB (Fig. 1). Ignoring or combining these poorly sampled bins does not affect the patterns of diversity, extinction or origination through time. Rarefaction analysis shows that the better-sampled time units, the Kopanskaya, Kzylsaiskaya and Staritskaya svitas (Fig. 1), may overestimate diversity by one, or at most two, families in comparison with the other time bins. Normalizing all time bin sizes to the range of 49–63 specimens cuts the diversity of the first three Triassic horizons by one or two families, hence making the PTB extinction seem larger (91% instead of 82% extinction rate) and depressing earliest Triassic diversity even more than has been indicated from the raw figures.

Scaling between local-scale or regional-scale observations such as these and the global scale is hard. Nonetheless, just as local-scale studies of marine PTB sections^{3,8} show patterns expected from global-scale studies^{3,20}, so this study, and similar investigations of the PTB in the Karoo basin of South Africa^{16,21}, indicate high familial extinction rates among tetrapods: 74% from a global database⁵, and 82% here. Our most striking finding has been that the high-diversity and complex latest Permian terrestrial ecosystems were volatile in terms of generic and familial turnover, but that when these ecosystems were largely destroyed by the PTB crisis the volatility disappeared, and recovery from low diversity was a slow

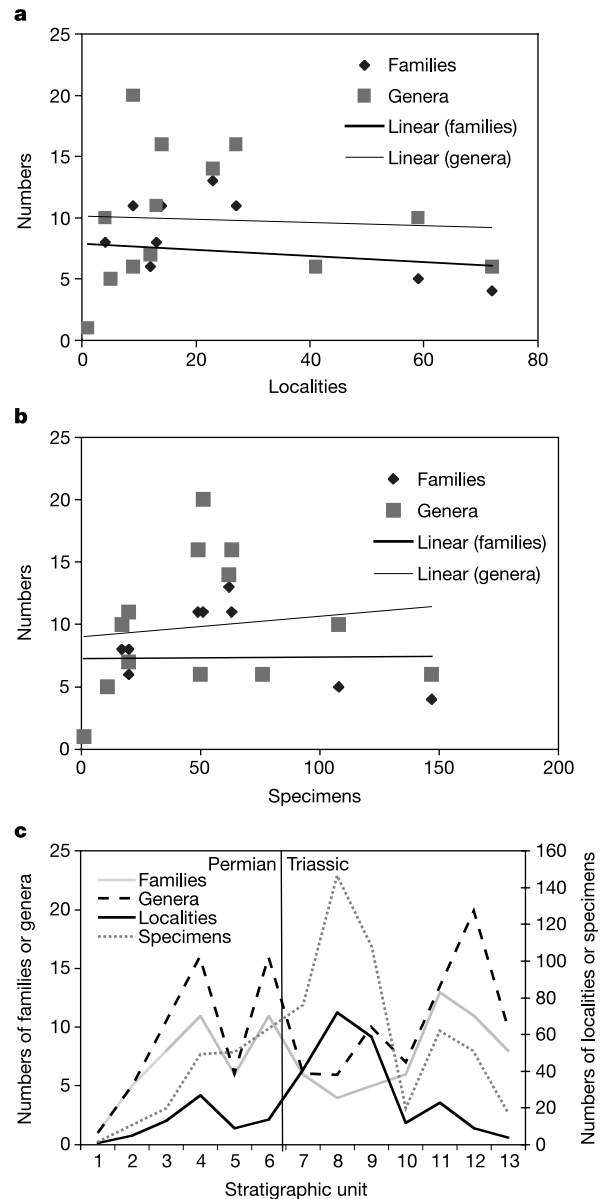


Figure 3 The data on tetrapod distributions from the South Urals are probably reliable, and cannot be accounted for simply by sampling (the patchiness of the rock and fossil record). **a, b**, Numbers of genera and families are not related to numbers of localities (**a**) or numbers of specimens (**b**). Best fitting straight lines show highly insignificant correlation coefficients. **c**, The distributions of generic and familial diversity through time (left-hand y-axis) follow similar curves, but these do not seem to relate to measures of sampling (numbers of localities and specimens per time bin, which themselves are correlated; right-hand y-axis). Stratigraphic units are as in Fig. 2.

process, with longer survivorship of genera and families and less turnover. Within the 15-Myr post-event window, full recovery of the ecosystems had not taken place. □

Methods

Database

The data set consists of records kept by V.P.T. at the Geological Institute of Saratov State University since the 1960s documenting every vertebrate specimen recovered in a broad geographic area about 400 km wide and 200 km long, bounded by Samara in the west and the Ural mountains in the east, and centring on the city of Orenburg. These records document a minimum of 675 specimens (isolated bones, complete skulls and complete skeletons) from 289 localities, each dated to one of the 13 time divisions of the Kazanian to Ladinian interval. The data are listed in refs 14 and 15. Lists of genera and families present in each time bin were compiled, together with the numbers of specimens of each taxon and

a note of the completeness of preservation (isolated bone, group of bones, complete skull, complete skeleton) for each.

In the past, Western authors have tended to rename Russian svitas as 'formations', and horizons as 'horizons', but this masks their true meanings. In Russia, horizons are the main regional stratigraphic units, identified primarily from their palaeontological characteristics, and they do not pertain to lithostratigraphic units. Svitas are largely lithostratigraphic units, given a locality name that is close to their characteristic exposure. The definition of a svita incorporates a mix of field lithological observations and biostratigraphic assumptions.

Analysis

The records were converted into range charts (Fig. 1), including Lazarus taxa²², from which total numbers (*N*) and numbers of originations (*O*) and extinctions (*E*) per time bin were calculated. Percentage origination and extinction metrics (*O/N*, *E/N*) were calculated for each time bin (Fig. 2). There are many other possible measures of extinction and origination rates, most calculated with respect to time; such measures would be inappropriate here because the durations of the svitas are poorly constrained. Boundary-crossing measures of extinction and origination rates were not used because the sample sizes are small, and 10 of the 38 families are restricted to one time bin and would have to be discarded. Generic rates are not presented because many genera are singletons (restricted to one time bin) and most are in need of taxonomic revision. Binomial error bars³⁰ are calculated for the percentage metrics.

The possible influence of sampling was assessed from the raw data (Fig. 3), and by the application of three sampling standardizations. In the first standardization, units that had yielded fewer than 50 specimens were ignored (namely the Osinovskaya, Belebey, Bolshekinelskaya, Gostevskaya and Bukobay svitas); sample sizes then ranged from 49 to 147 specimens. In the second sampling standardization, the two oldest units were ignored, and the others with small sample sizes were combined with adjacent units (Bolshekinelskaya + Amanakskaya, Gostevskaya + Petropavlovskaya, Donguz + Bukobay), yielding a range of sample sizes from 50 to 147 specimens. In the third sampling standardization, rarefaction analysis was applied to the units that had yielded larger samples of specimens (Kopanskaya, Kzylsaiskaya, Staritskaya) to assess what their apparent diversity would have been had the sample size been 50, within the range 49–63 specimens, as for the other moderately well sampled units.

The data sets and analyses are available as Supplementary Data, and may be downloaded at <http://palaeo.gly.bris.ac.uk/Data/RussiaPTr.xls>.

Dating

The timescale indicated in Fig. 1 is based on refs 10, 23, 24 and 25. The date for the Permian–Triassic boundary, 251 Myr, from ref. 10, has been debated²⁶, but is widely accepted and will be the accepted date in the new Cambridge geologic timescale^{27,28}. Other aspects of the scales may seem less familiar, in that the Kazanian and Tatarian are much longer than is often assumed, 16 Myr instead of 4–5 Myr, and the Middle Triassic is dated as older than normally accepted. Should the old dates prove to be correct, and the newer ones incorrect, the conclusions here are not affected because we do not make claims about the longer-term timing of events, nor do we present rates of origination or extinction calculated against time.

Received 2 July; accepted 18 August 2004; doi:10.1038/nature02950.

- Erwin, D. H. The Permo-Triassic extinction. *Nature* **367**, 231–236 (1994).
- Benton, M. J. *When Life Nearly Died* (Thames & Hudson, London, 2003).
- Jin, Y. G. *et al.* Pattern of marine mass extinction near the Permian–Triassic boundary in south China. *Science* **289**, 432–436 (2000).
- Sepkoski, J. J. Jr in *Global Events and Event Stratigraphy* (ed. Walliser, O. H.) 35–52 (Springer, Berlin, 1996).
- Maxwell, W. D. Permian and Early Triassic extinction of nonmarine tetrapods. *Palaeontology* **35**, 571–583 (1992).
- Raup, D. M. Size of the Permo-Triassic bottleneck and its evolutionary implications. *Science* **206**, 217–218 (1979).
- McKinney, M. L. Extinction selectivity among lower taxa—gradational patterns and rarefaction error in extinction estimates. *Paleobiology* **21**, 300–313 (1995).
- Twitchett, R. J., Looy, C. V., Morante, R., Visscher, H. & Wignall, P. B. Rapid and synchronous collapse of marine and terrestrial ecosystems during the end-Permian biotic crisis. *Geology* **29**, 351–354 (2001).
- Benton, M. J. & Twitchett, R. J. How to kill (almost) all life: the end-Permian extinction event. *Trends Ecol. Evol.* **18**, 358–365 (2003).
- Bowring, S. A. *et al.* U/Pb zircon geochronology of the end-Permian mass extinction. *Science* **280**, 1039–1045 (1998).
- Wignall, P. B. & Twitchett, R. J. Oceanic anoxia and the end Permian mass extinction. *Science* **272**, 1155–1158 (1996).
- Berner, R. A. Examination of hypotheses for the Permo-Triassic boundary extinction by carbon cycle modeling. *Proc. Natl Acad. Sci. USA* **99**, 4172–4177 (2002).
- Newell, A. J., Tverdokhlebov, V. P. & Benton, M. J. Interplay of tectonics and climate on a transverse fluvial system, Upper Permian, Southern Uralian foreland basin, Russia. *Sediment. Geol.* **127**, 11–29 (1999).
- Tverdokhlebov, V. P., Tverdokhlebova, G. I., Surkov, M. V. & Benton, M. J. Tetrapod localities from the Triassic of the SE of European Russia. *Earth Sci. Rev.* **60**, 1–66 (2003).
- Tverdokhlebov, V. P., Tverdokhlebova, G. I., Minikh, A. V., Surkov, M. V. & Benton, M. J. Upper Permian vertebrates and their sedimentological context in the South Urals, Russia. *Earth Sci. Rev.* (in the press) (2004).
- Ward, P. D., Montgomery, D. R. & Smith, R. Altered river morphology in South Africa related to the Permian–Triassic extinction. *Science* **289**, 1740–1743 (2000).
- Michaelsen, P. Mass extinction of peat-forming plants and the effect on fluvial styles across the

- Permian–Triassic boundary, Northern Bowen Basin, Australia. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **179**, 173–188 (2002).
- Peters, S. E. & Foote, M. Biodiversity in the Phanerozoic: a reinterpretation. *Paleobiology* **27**, 583–601 (2001).
- Smith, A. B. Large-scale heterogeneity of the fossil record: implications for Phanerozoic biodiversity studies. *Phil. Trans. R. Soc. Lond. B* **356**, 351–367 (2001).
- Benton, M. J. Diversification and extinction in the history of life. *Science* **268**, 52–58 (1995).
- Smith, R. M. H. & Ward, P. D. Pattern of vertebrate extinctions across an event bed at the Permian–Triassic boundary in the Karoo Basin of South Africa. *Geology* **29**, 1147–1150 (2001).
- Jablonski, D. in *Dynamics of Extinction* (ed. Elliott, D. K.) 183–230 (Wiley, New York, 1986).
- Cassinis, G., Nicosia, U., Lozovsky, V. R. & Gubin, Y. M. A view of the Permian continental stratigraphy of the Southern Alps, Italy, and general correlation with the Permian of Russia. *Permian* **40**, 4–16 (2002).
- Kozur, H. W. Integrated ammonoid, conodont and radiolarian zonation of the Triassic and some remarks to stage/substage subdivision and the numeric age of the Triassic stages. *Albertiana* **28**, 570–574 (2003).
- Jin, Y., Wardlaw, B. R., Glenister, B. F. & Kotlyar, G. V. Permian chronostratigraphic divisions. *Episodes* **20**, 10–15 (1997).
- Mundil, R. *et al.* Timing of the Permian–Triassic biotic crisis: implications from new zircon U/Pb age data (and their limitations). *Earth Planet. Sci. Lett.* **187**, 131–145 (2001).
- Ogg, J. G. Status of divisions of the International Geologic Time Scale. *Lethaia* **37**, 183–199 (2004).
- Gradstein, F. M., Ogg, J. G. & Smith, A. G. (eds) *A Geologic Time Scale 2004* (Cambridge Univ. Press, in the press).
- Benton, M. J. Testing the roles of competition and expansion in tetrapod evolution. *Proc. R. Soc. Lond. B* **263**, 641–646 (1996).
- Raup, D. M. in *Analytical Paleobiology, Short Courses in Paleontology No. 4* (eds Gilinsky, N. L. & Signor, P. W.) 207–216 (Paleontological Society, Knoxville, Tennessee, 1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Donoghue, S. Braddy, H. Falcon-Lang and R. Twitchett for helpful comments on the manuscript, and S. Powell for Fig. 1. Funding for this work was provided by the Royal Society, INTAS, RFBR, Russian Ministry of Education, the Leverhulme Trust, and the National Geographic Society.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.B. (mike.benton@bris.ac.uk).

Long-term decline in krill stock and increase in salps within the Southern Ocean

Angus Atkinson¹, Volker Siegel², Evgeny Pakhomov^{3,4} & Peter Rothery⁵

¹British Antarctic Survey, Natural Environment Research Council, High Cross, Madingley Road, Cambridge CB3 0ET, UK

²Sea Fisheries Institute, Palmalle 9, 22767 Hamburg, Germany

³Department of Earth and Ocean Sciences, University of British Columbia, 6339 Stores Rd, Vancouver, British Columbia, V6T 1Z4, Canada

⁴Department of Zoology, Faculty of Science and Technology, University of Fort Hare, Private Bag X1314, Alice 5700, South Africa

⁵NERC Centre for Ecology and Hydrology, CEH Monks Wood, Abbots Ripton, Huntingdon PE28 2LS, UK

Antarctic krill (*Euphausia superba*) and salps (mainly *Salpa thompsoni*) are major grazers in the Southern Ocean^{1–4}, and krill support commercial fisheries⁵. Their density distributions^{1,3,4,6} have been described in the period 1926–51, while recent localized studies^{7–10} suggest short-term changes. To examine spatial and temporal changes over larger scales, we have combined all available scientific net sampling data from 1926 to 2003. This database shows that the productive southwest Atlantic sector contains >50% of Southern Ocean krill stocks, but here their density has declined since the 1970s. Spatially, within their habitat, summer krill density correlates positively with chlorophyll concentrations. Temporally, within the southwest Atlantic, summer krill densities correlate positively with sea-ice extent the

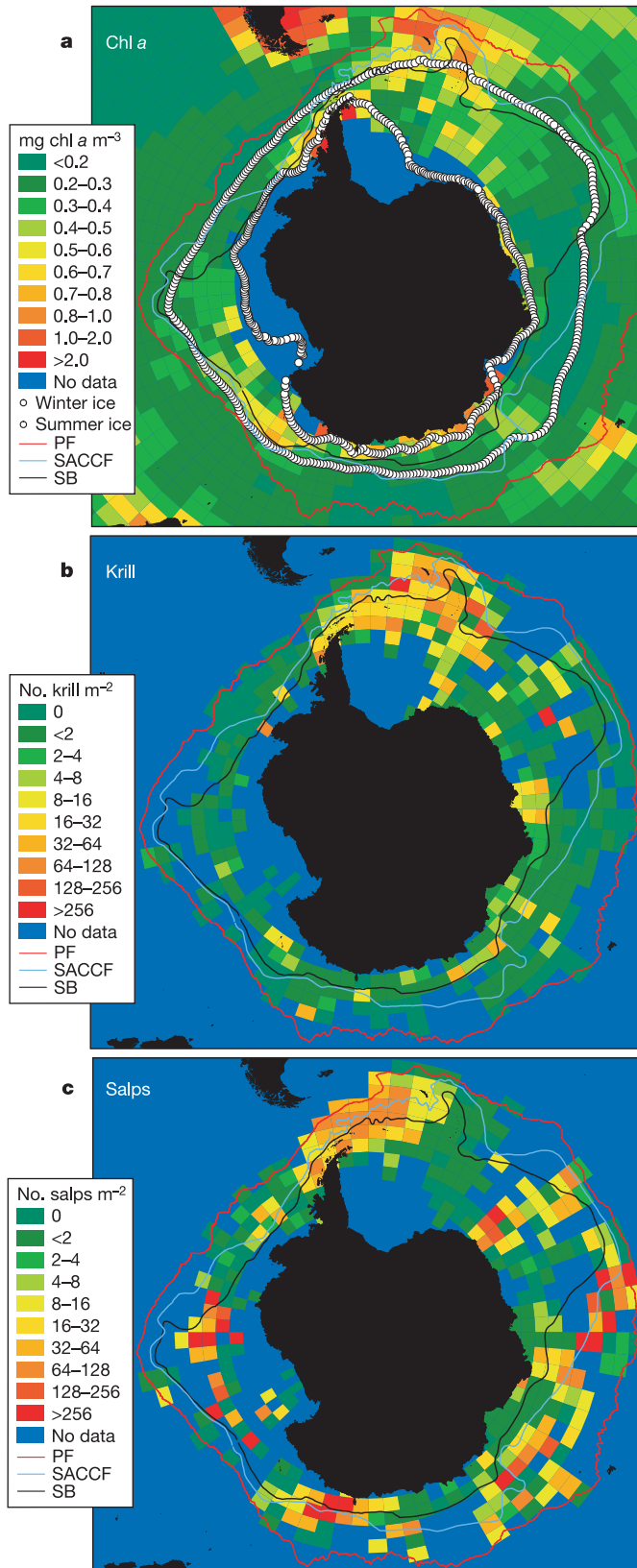


Figure 1 Krill, salps and their food. **a**, Mean (November–April) chlorophyll *a* (chl *a*) concentration, 1997–2003. **b**, Mean krill density (6,675 stations, 1926–2003). **c**, Mean salp density (5,030 stations, 1926–2003). $\log_{10}(\text{no. krill m}^{-2}) = 1.2 \log_{10}(\text{mg chl } a \text{ m}^{-3}) + 0.83$ ($R^2 = 0.051$, $P = 0.017$, $n = 110$ grid cells). Historical mean positions are shown for the PF²⁹, Southern ACC Front (SACCF)³⁰, SB³⁰ and northern 15% sea-ice concentrations in February and September (1979–2004 means).

previous winter. Summer food and the extent of winter sea ice are thus key factors in the high krill densities observed in the southwest Atlantic Ocean. Krill need the summer phytoplankton blooms of this sector, where winters of extensive sea ice mean plentiful winter food from ice algae, promoting larval recruitment^{7–11} and replenishing the stock. Salps, by contrast, occupy the extensive lower-productivity regions of the Southern Ocean and tolerate warmer water than krill^{2–4,12}. As krill densities decreased last century, salps appear to have increased in the southern part of their range. These changes have had profound effects within the Southern Ocean food web^{10,13}.

Our database comprises 11,978 net hauls from 9 countries, spanning the summers of 1926–39 and 1976–2003. This database shows a concentration of krill in the productive southwest (SW) Atlantic sector (Fig. 1a, b). On the basis of catches with equivalent nets (Supplementary Information), 58–71% of krill are located here.

Potential krill habitat lies between the Polar Front (PF) to the north and the ice-covered Antarctic shelves to the south^{1,7,14}, but krill do not occupy all of this range. There are various explanations for regional variations in krill density, invoking the seasonal ice

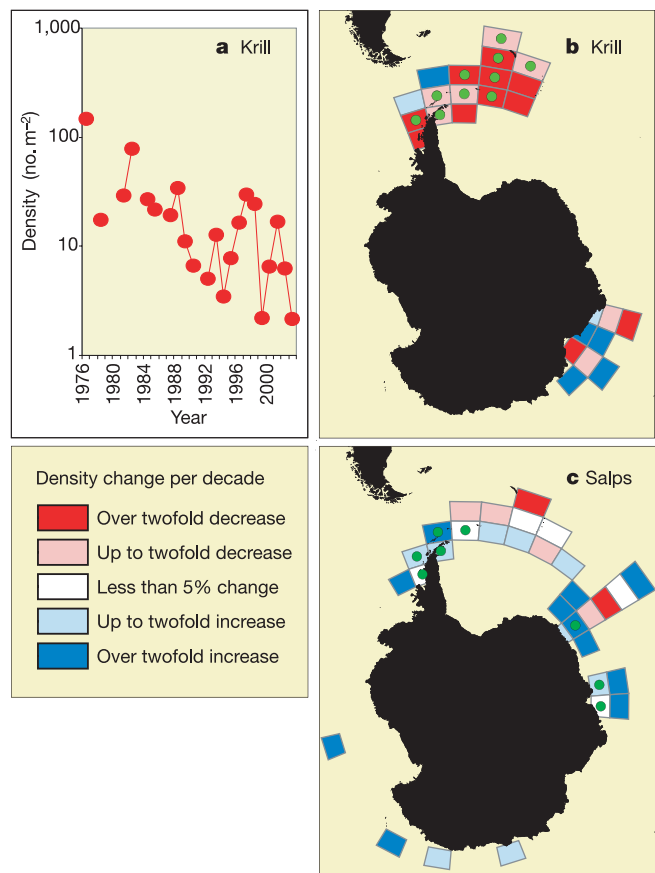


Figure 2 Temporal change of krill and salps. **a**, Krill density in the SW Atlantic sector (4,948 stations in years with >50 stations). Temporal trends include **b**, post-1976 krill data from scientific trawls; **c**, 1926–2003 circumpolar salp data south of the SB. Regressions of $\log_{10}(\text{mean no. m}^{-2})$ on year were calculated for cells with ≥ 3 yr of data, weighted by number of stations in that year. One-sample *t*-tests supported a post-1976 decrease in krill density in the SW Atlantic (scientific trawls: $t = -3.4$, $P = 0.004$, 16 cells, smaller nets: $t = -2.5$, $P = 0.04$, 8 cells). Salp densities increased south of the SB after 1926 ($t = 3.1$, $P = 0.004$, 32 cells) Green spots denote cells usable in the spatio-temporal model.

Table 1 Significant temporal trends krill and salp density

Taxon	Subset of data analysed			Estimated % increase (+) or decrease (–) per decade (s.e.m)		P value
	Era	Net type	Region			
Salps	Post 1926 (34 yr)	All	Circumpolar (8 grid cells)	+66 (23)	+ 87 (20)	0.007 < 0.001
Krill	Post 1976 (21 yr)	Scientific trawls	SW Atlantic (10 grid cells)	–38 (24)	– 38 (15)	0.12 0.023
Krill	Post 1976 (14 yr)	All other nets	SW Atlantic (8 grid cells)	–75 (21)	– 81 (16)	0.004 < 0.001

A spatio-temporal model for $\log_{10}$ (no. krill or salps m^{-2}) with linear trend, grid cell effects, and including random year effects (upright font) or ignoring random year effects (italic font). See Methods.

zone^{12,15} (SIZ), the Southern Boundary (SB) of the Antarctic Circumpolar Current¹⁶ (ACC) or the province to its south¹². But none of these relationships hold at the circumpolar scale¹⁷. For example, krill densities are high near South Georgia, north of both the SB and the SIZ (Fig. 1). However, all of these models link krill density to the abundance of food, which is likely to be the primary factor determining abundance. Together, sea ice, oceanography and nutrients promote primary production near shelves, fronts and ice edges¹⁸, and krill occupy this full range of habitats. Within the distributional range of krill, their mean density correlates positively with the concentration of chlorophyll *a* (chl *a*; Fig. 1).

Salps tend to occupy warmer water than krill^{2–4,12}, and prefer oceanic regions with lower food concentrations^{2,3}. Thus the lower productivity across most of the ACC means that the habitat of salps is much larger than that of krill, with no concentration into one sector (Fig. 1c). The hotspot of krill in the SW Atlantic—a feature very unlike that of zooplankton⁶—suggests an ability to maintain their 5–7-yr life cycle here, withstanding entrainment into the great current systems encircling Antarctica.

We studied temporal trends in krill and salp density using a grid, which incorporates inter-annual variability in density (Fig. 2a) into a time-series regression within each of its cells. These are used in one-sample *t*-tests (for example, Fig. 2b, c) and a spatio-temporal model (Table 1). Salp densities increased south of the SB over the whole time series. For krill in the SW Atlantic sector, densities have declined significantly since 1976.

Monitoring surveys^{7–10} show shorter-term changes in krill and salp density. However, these are too localized to tell whether they reflect changes in overall population size¹⁷. The trends reported here are longer-term and over larger scales, supported by independent surveys with scientific trawls (that is, Rectangular midwater

trawls and Isaac Kidd midwater trawls) and with smaller nets (Supplementary Information).

Controls on grazer populations include top-down (predation) and bottom-up (resource-based) factors¹⁷. We have examined temporal links between krill density and a key physical parameter—winter sea ice^{7–11}. To find the appropriate scale for analysis, we compared krill densities between the eastern (30°–50° W) and western (50°–70° W) sub-areas of the SW Atlantic. Over the 19 available summers, krill density in the east and west are positively related ($R^2 = 0.47$, $P < 0.001$). Whether this reflects advection^{19,20} or common controls on krill populations¹⁷, it suggests a basin-scale synchrony in the inter-annual signal.

We therefore compared net data from the whole SW Atlantic to indices of sea ice. Here, summer krill density correlates to both the duration (Fig. 3a) and the extent (Fig. 3b) of sea ice the previous winter in the same area. Salps showed no such relationships, despite a negative one being postulated previously off the Antarctic Peninsula⁸. With shorter life cycles than krill and explosive population growth rates, salps can respond to environmental variation over shorter timescales^{2–4}.

The population size of krill has been linked both to predation controls^{11,13,17,20,21} and to food resources within winter sea ice^{7–11}. Our results are, to our knowledge, the first to show a direct, large-scale link between annual krill density and sea-ice cover. This is not a localized, short-term effect—it relates to >50% of their stock and the data span 1926–2003. Given the problems of net sampling⁵, the existence of a relationship with sea ice suggests that this factor plays a dominant role, not just in krill recruitment^{7–11}, but also in their population size.

Various mechanisms have been proposed to explain how sea ice benefits krill^{7–11,22}. Sea-ice algae are a critical food resource, boosting early adult spawning in spring or survival of the larvae the following winter. Sea ice could also shield krill from predation. However, the only link with summer krill density that we found was with ice cover the preceding winter—ice cover from the winter before that had no significant effect. A 6-month lag points to larval over-wintering as a key process affected by ice—larvae need to survive their first winter to recruit the following summer, and thus replenish the adult population. At the Antarctic Peninsula, only a few years of strong recruitment per decade are needed to maintain the local krill population^{7–11}.

These positive temporal relationships between krill density and sea-ice extent (Fig. 3) do not equate to positive spatial relationships (Fig. 1a, b). High krill densities can be found at South Georgia outside the SIZ, whereas low densities coincide with an extensive SIZ, for example, in the SW Indian sector. Hence sufficient winter ice in the major spawning and nursery areas (the Antarctic Peninsula and Southern Scotia Arc^{1,7,11,23}) affects krill density across a whole ocean basin, including areas north of the SIZ. Krill larvae do not merely need to survive the winter—they need to double in length¹¹. Sufficient food year-round may thus be the feature of the SW Atlantic that maintains its high krill stocks.

The western Antarctic Peninsula is one of the world's fastest warming areas, and (atypically for the Southern Ocean) winter sea-ice duration in this sector is shortening²⁴. Key spawning and nursery

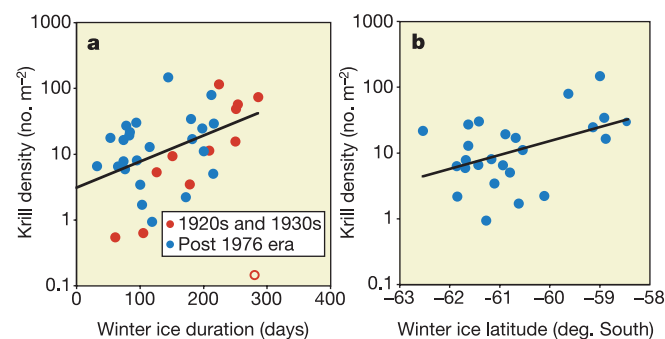


Figure 3 Krill–ice relationships. Annual mean density of krill across the SW Atlantic versus **a**, sea-ice duration²⁷ (that is, days of fast ice observed at the South Orkneys the previous winter), and **b**, the mean September latitude of 15% ice cover along a transect¹⁰ across the western Scotia Sea. Regression identified one outlier season (1934, open circle) with exceptionally long ice duration and only 24 net stations, so for the remaining years $\log_{10}(\text{no. krill } m^{-2}) = 0.49 + 0.0040 (\text{sea-ice duration, days})$ $R^2 = 0.21$, $P = 0.006$, $n = 35$. $\log_{10}(\text{no. krill } m^{-2}) = 14 + 0.21 (\text{sea ice latitude, degrees})$ $R^2 = 0.21$, $P = 0.02$, $n = 25$.

areas of krill are thus located in a region that is sensitive to environmental change. Deep ocean temperatures have increased²⁵, and a circumpolar, pre 1970s decrease in sea ice²⁶ has been indicated at several locations^{27,28}. The regional decrease in a high-latitude species with high food requirements (krill) coincides with an increase in a lower-latitude group with lower food requirements (salps). However, as the mechanisms underlying these changes are uncertain, future predictions must be cautious.

These changes among key species have profound implications for the Southern Ocean food web. Penguins, albatrosses, seals and whales have wide foraging ranges but are prone to krill shortage^{3,10,13,21}. Thus the wide extent of our indicated change in krill density—not just its magnitude—is important. The basin-scale decline in krill may underlie the post 1980s shift in demography of krill predators, seen across the SW Atlantic^{10,13}. Earlier last century, over-exploitation of whales preceded a rapid increase in smaller krill predators such as fur seals in the SW Atlantic^{20,21}. Added to this shift in the predator balance, a return of the whales to pre-exploitation levels now faces the further problem of lower krill density. □

Methods

Satellite data

Average values of SeaWiFS Level-3 standard mapped images of chl *a* concentration were calculated in Arc GIS 8.2 for grid cells with data for November–April (Fig. 1a). Sea-ice images, calculated from DMSP-SSM passive microwave data by NOAA/NCEP, were pre-viewed as the northern extent of 15% ice concentration to remove spurious values before calculating mean monthly positions.

The krill and salp database

The full database (Supplementary Table 2) comprises data from the UK, Germany, USA, the Ukraine, South Africa, Japan, Australia, Poland and Spain. All are non-targeted oblique or vertical hauls from pre-fixed positions. The data were either from samples sorted by the authors, available within our institutes, sent by our collaborators, or transcribed from the literature. Krill densities (no. m⁻²) include only postlarvae, or just the krill >19 mm long from the *Discovery* (1926–51) era. Salp densities are the total solitary plus aggregate individuals of *S. thompsoni* and the rarer *Ihleia racovitzai*, pooled to avoid identification problems. Historical *Discovery* data (1926–51) were cross-checked from three archived sources: the net sampling logs, the original tables used to construct the published figures^{1,4}, and an electronic krill database. The *Discovery Report Station Lists*^{1,4} were used to calculate densities.

Extraction and analysis of data

From the full database, we extracted November–April data south of the PF where at least the topmost 75 m (krill) or 100 m (salps) was sampled. Sampling was mainly much deeper (Supplementary Table 3) and 90% was in December–March. Regression of krill density on chl *a* concentration (Fig. 1) used mean grid cell values, and was restricted to post-1976 scientific trawl data in cells where krill were caught.

We tested spatio-temporal trends among: stations south of the PF, PF to SB, south of the SB, the post 1926 era, the post 1976 era, the SW Atlantic sector only and circumpolar. We report only trends significant in two tests—first, a one-sample *t*-test of whether the regression slopes of cells (Fig. 2) differ from zero, supporting a widespread shift in abundance. Second, to further test for a temporal trend we used a spatio-temporal model: $y_{kt} = g_k + bt + B_t + E_{kt}$, where y_{kt} is the log₁₀ transformed mean density (+*c*) in grid square *k* and season *t*, g_k is a fixed effect for grid square *k*, *b* is the slope, the average change per season on a log₁₀ scale, B_t is a random effect for season *t* and E_{kt} is a random square by season effect. The constant *c* (added to allow for log transformation of zero densities) was half the minimum density. The variance of the cell-season effect was assumed to be inversely proportional to the number of net hauls (n_{kt}) for each cell-season combination. Cells with sufficient data for inclusion in the mixed model (Fig. 2) were defined as those with at least 5 seasons and 50 stations. The model was fitted by residual maximum likelihood, with the *t*-test for trend based on $n - 2$ degrees of freedom, where *n* is the number of seasons. A general linear model ignoring the random year effect (that is, the values shown in italics in Table 1), gave similar estimated slopes but smaller standard errors—an effect of pseudo-replication from treating the observations as statistically independent.

Potential sampling artefacts

The asymmetrical circumpolar distribution of krill density persisted last century and is supported independently by differing sampling gear (Supplementary Fig. 4). Our spatio-temporal analyses encompass broad-scale spatial differences in sampling location, but Supplementary Fig. 5 shows that at finer scales, sampling has focused increasingly onto productive shelf/shelf break areas favoured by krill^{14,17} but not by salps^{2,3}. The trends are thus not artefacts of sampling emphasis. Likewise, any systematic changes in sampling method (Supplementary Table 3) observed would tend to increase krill density in recent years rather than decrease it. Station positions were pre-fixed, so generally sampled on arrival—that is, at random times of the day or night, thus not biasing our grid-based

analyses. The sea-ice relationships were supported by multiple subsets of krill data and sea-ice indices (Supplementary Table 4).

Received 17 May; accepted 7 September 2004; doi:10.1038/nature02996.

- Marr, J. W. S. The natural history and geography of the Antarctic krill (*Euphausia superba* Dana). *Discovery Rep.* **32**, 33–464 (1962).
- Le Fèvre, J., Legendre, L. & Rivkin, R. B. Fluxes of biogenic carbon in the Southern Ocean: roles of large microphagous zooplankton. *J. Mar. Syst.* **17**, 325–345 (1998).
- Pakhomov, E. A., Froneman, P. W. & Perissinotto, R. Salp/krill interactions in the Southern Ocean: spatial segregation and implications for the carbon flux. *Deep-Sea Res. II* **49**, 1881–1907 (2002).
- Foxton, P. The distribution and life history of *Salpa thompsoni* Foxton with observations on a related species, *Salpa gerlachei*. *Discovery Rep.* **34**, 1–116 (1966).
- Everson, E. *Krill. Biology, Ecology and Fisheries* (Blackwell Science, London, 2000).
- Baker, A. de C. The circumpolar continuity of Antarctic plankton species. *Discovery Rep.* **27**, 201–218 (1954).
- Siegel, V. Krill (Euphausiacea) demography and variability in abundance and distribution. *Can. J. Fish. Aquat. Sci.* **57**, 151–167 (2000).
- Loeb, V. et al. Effects of sea-ice extent and krill or salp dominance on the Antarctic food web. *Nature* **387**, 897–900 (1997).
- Quetin, L. B. & Ross, R. M. Episodic recruitment in Antarctic krill *Euphausia superba* in the Palmer LTER study region. *Mar. Ecol. Prog. Ser.* **259**, 185–200 (2003).
- Fraser, W. R. & Hofmann, E. E. A predator's perspective on causal links between climate change, physical forcing and ecosystem response. *Mar. Ecol. Prog. Ser.* **265**, 1–15 (2003).
- Quetin, L. B., Ross, R. M., Frazer, T. K. & Haberman, K. L. Factors affecting distribution and abundance of zooplankton, with an emphasis on Antarctic krill, *Euphausia superba*. *Antarct. Res. Ser.* **70**, 357–371 (1996).
- Nicol, S. et al. Ocean circulation off east Antarctica affects ecosystem structure and sea-ice extent. *Nature* **406**, 504–507 (2000).
- Reid, K. & Croxall, J. P. Environmental response of upper trophic-level predators reveals a system change in an Antarctic marine ecosystem. *Proc. R. Soc. Lond. B* **268**, 377–384 (2001).
- Nicol, S., Constable, A. & Pauly, T. Estimates of circumpolar Antarctic krill abundance based on recent acoustic density measurements. *CCAMLR Sci.* **7**, 87–99 (2000).
- Brierley, A. S. et al. Antarctic krill under sea ice: elevated abundance in a narrow band just south of the ice edge. *Science* **295**, 1890–1892 (2002).
- Tynan, C. Ecological importance of the Southern Boundary of the Antarctic Circumpolar Current. *Nature* **392**, 708–710 (1998).
- Constable, A. J., Nicol, S. & Strutton, P. G. Southern Ocean productivity in relation to spatial and temporal variation in the physical environment. *J. Geophys. Res.* **108**, doi: 10.1029/2001JC001270 (2003).
- Arrigo, K. R., Worthen, D., Schnell, A. & Lizotte, M. P. Primary production in Southern Ocean waters. *J. Geophys. Res.* **103**, 15587–15600 (1998).
- Hofmann, E. E., Klinck, J. M., Locarnini, R. A., Fach, B. & Murphy, E. Krill transport in the Scotia Sea and environs. *Antarct. Sci.* **10**, 406–415 (1998).
- Murphy, E. J. Spatial structure of the Southern Ocean ecosystem: predator-prey linkages in Southern Ocean food webs. *J. Anim. Ecol.* **64**, 333–347 (1995).
- Laws, R. M. The ecology of the Southern Ocean. *Am. Sci.* **73**, 26–40 (1985).
- Smetacek, V., Scharek, R. & Nöthig, E. M. in *Antarctic Ecosystems. Ecological Change and Conservation* (eds Kerry, K. R. & Hempel, G.) 103–114 (Springer, Berlin, 1990).
- Hofmann, E. E. & Hüsrevoglu, Y. S. A circumpolar modelling study of habitat control of Antarctic krill (*Euphausia superba*) reproductive success. *Deep-Sea Res. II* **50**, 3121–3142 (2003).
- Parkinson, C. L. Trends in the length of the Southern Ocean sea-ice season, 1979–99. *Ann. Glaciol.* **34**, 435–440 (2002).
- Gille, S. T. Warming of the Southern Ocean since the 1950s. *Science* **295**, 1275–1277 (2002).
- de la Mare, W. K. Abrupt mid-twentieth-century decline in Antarctic sea ice from whaling records. *Nature* **389**, 57–60 (1997).
- Clarke, A. & Harris, C. M. Polar marine ecosystems: major threats and future change. *Environ. Conserv.* **30**, 1–25 (2003).
- Curran, M. A. J., van Ommen, T. D., Morgan, V. L., Phillips, K. L. & Palmer, A. S. Ice core evidence for Antarctic sea ice decline since the 1950s. *Science* **302**, 1203–1206 (2003).
- Moore, J. K., Abbott, M. R. & Richman, J. G. Location and dynamics of the Antarctic Polar front from satellite sea surface temperature data. *J. Geophys. Res.* **104**, 3059–3073 (1999).
- Orsi, A. H., Whitworth, T. III & Nowlin, W. D. J. On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep-Sea Res. I* **42**, 641–673 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank V. Loeb and R. Hewitt for sharing the extensive data from the AMLR Program; A. Clarke, G. Hosie, S. Chiba, J. Nishikawa, R. Anadón, P. Ward and A. de C. Baker for providing further data and information; P. Fretwell, A. Fleming, S. Grant and S. Thorpe for data handling; and A. Hirst, A. Clarke, P. Ward, K. Schmidt, D. Pond, P. Trathan, G. Tarling and E. Murphy for comments. Satellite data were provided courtesy of NASA GSFC, the SeaWiFS Project and NOAA. The Department of Environmental Affairs & Tourism (Pretoria, South Africa) provided funds and facilities for South African collections, through the South African National Antarctic Program. A.A.'s contribution was funded under the British Antarctic Survey's DYNAMOE Programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.A. (aat@bas.ac.uk).

Adaptive divergence in pigment composition promotes phytoplankton biodiversity

Maayke Stomp^{1*}, Jef Huisman^{1*}, Floris de Jongh¹, Annelies J. Veraart¹, Daan Gerla¹, Machteld Rijkeboer², Bas W. Ibelings², Ute I. A. Wollenzien³ & Lucas J. Stal³

¹Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands

²Netherlands Institute of Ecology (NIOO-KNAW), Centre for Limnology, Rijksstraatweg 6, 3631 AC Nieuwersluis, The Netherlands

³Netherlands Institute of Ecology (NIOO-KNAW), Centre for Estuarine and Marine Ecology, PO Box 140, 4400 AC Yerseke, The Netherlands

* These authors contributed equally to this work

The dazzling diversity of the phytoplankton has puzzled biologists for decades^{1–5}. The puzzle has been enlarged rather than solved by the progressive discovery of new phototrophic microorganisms in the oceans, including picocyanobacteria^{6,7}, pico-eukaryotes⁸, and bacteriochlorophyll-based^{9–11} and rhodospin-based phototrophic bacteria^{12,13}. Physiological and genomic studies suggest that natural selection promotes niche differentiation among these phototrophic microorganisms, particularly with respect to their photosynthetic characteristics^{14–16}. We have analysed competition for light between two closely related picocyanobacteria of the *Synechococcus* group that we isolated from the Baltic Sea¹⁷. One of these two has a red colour because it contains the pigment phycoerythrin, whereas the other is blue-green because it contains high contents of the pigment phycocyanin. Here we report theory and competition experiments that reveal stable coexistence of the two picocyanobacteria, owing to partitioning of the light spectrum. Further competition experiments with a third marine cyanobacterium, capable of adapting its pigment composition, show that this species persists by investing in the pigment that absorbs the colour not used by its competitors. These results demonstrate the adaptive significance of divergence in pigment composition of phototrophic microorganisms, which allows an efficient utilization of light energy and favours species coexistence.

Phytoplankton harvest light with photosynthetic pigments^{18,19}. These pigments absorb photons in specific regions of the light spectrum, while reflecting or scattering photons in other regions of the spectrum. The latter determines the colour of the pigment. The combination of pigments in a species determines which part of the light spectrum the species can utilize for photosynthesis. Figure 1a and c shows chemostat cultures of two *Synechococcus*-type picocyanobacteria, called BS4 and BS5, that we isolated from 10 m depth during a cruise on the Baltic Sea in August 1996. The two picocyanobacteria are genetically very similar, with less than 1% sequence divergence in their ITS-1 sequences¹⁷, yet differ remarkably in colour. The blue-green colour of BS4 is a result of the pigment phycocyanin, which absorbs photons in the orange-red part of the spectrum (620–630 nm; Fig. 1b). The red colour of BS5 is due to the pigment phycoerythrin, which absorbs photons in the green-yellow part of the spectrum (560–570 nm; Fig. 1d). The absorption spectra of BS4 and BS5 both show additional peaks in the blue and the red part of the spectrum (at ~430 and ~680 nm, respectively), caused by the pigment chlorophyll *a*, shared by all oxygenic phototrophic organisms.

Would the differences in pigment composition allow coexistence of the two picocyanobacteria, through a subtle form of niche differentiation? Existing theory and experiments on phytoplankton

competition for light predict competitive exclusion^{20–22}. That is, photons absorbed by one species are not available for photosynthesis by other species. As a result, species interact by shading each other, and only the strongest competitor for light survives. However, this previous work neglected the spectral aspects of light. Here, we develop a competition model that does include the light spectrum. Consider a mixture of *n* phytoplankton species. Let *N_i* denote the population density (in biovolume ml^{–1}) of phytoplankton species *i*, and let *I*(λ, *z*) denote the light intensity (in μmol photons m^{–2} s^{–1}) of wavelength λ at depth *z*. According to the Lambert-Beer law, the underwater light spectrum can be described as:

$$I(\lambda, z) = I_{in}(\lambda) \exp \left(- \sum_{i=1}^n k_i(\lambda) N_i z - K_{bg}(\lambda) z \right) \quad (1)$$

where *I_{in}*(λ) is the spectrum of the incident light intensity, *k_i*(λ) is the specific light absorption spectrum of phytoplankton species *i*, and *K_{bg}*(λ) is the background light absorption spectrum.

The number of photons absorbed over the photosynthetically active range of 400–700 nm by a single phytoplankton of species *i* at depth *z*, which will be denoted as γ_{*i*}(*z*), can be quantified as²³:

$$\gamma_i(z) = \int_{400}^{700} I(\lambda, z) k_i(\lambda) d\lambda \quad (2)$$

In photosynthesis, one absorbed photon can excite at most one electron. Hence, the number of photons absorbed by species *i* is the relevant quantity for photosynthesis, and thus for population growth of species *i*. Accordingly, if all populations are uniformly distributed over the surface mixed layer, the population dynamics of *n* competing species in the surface mixed layer can be described as:

$$\frac{dN_i}{dt} = \frac{\phi_i}{z_m} \int_0^{z_m} \gamma_i(z) N_i dz - L_i N_i \quad i = 1, \dots, n \quad (3)$$

where *z_m* is the mixing depth, φ_{*i*} is the photosynthetic efficiency of species *i* (that is, the efficiency with which absorbed photons are used for population growth), and *L_i* is the specific loss rate of species *i* due to factors like grazing, viruses, or other forms of cell death. We note that these differential equations are coupled via equation (1).

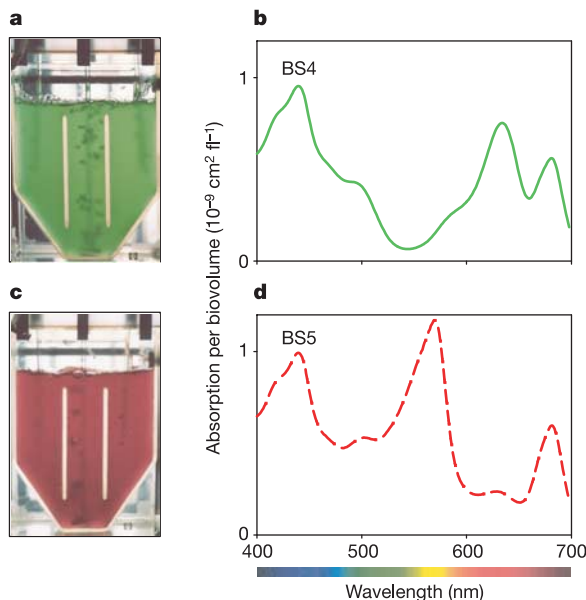


Figure 1 Optical characteristics of the picocyanobacteria BS4 and BS5. Monocultures of BS4 (a) and BS5 (c) grown in chemostats, and the light absorption spectra of BS4 (b) and BS5 (d).

Thus, the species interact by shading each other in specific regions of the underwater light spectrum.

To test the competition model defined by equations (1)–(3), we ran various experiments with BS4 and BS5 in light-limited chemostats. First, we used monoculture experiments under white light to estimate the model parameters of the two species (see Methods section). The model predictions fitted very well to the monoculture

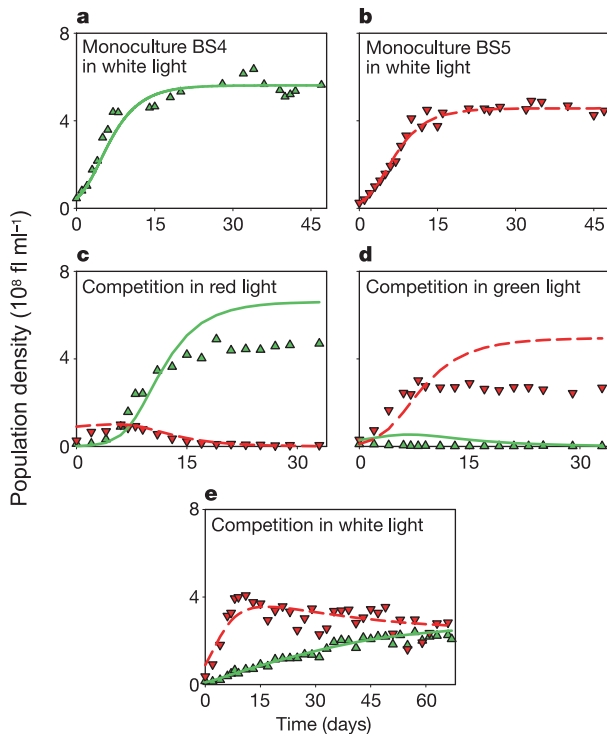


Figure 2 Monoculture experiments and competition experiments with the picocyanobacteria BS4 and BS5. Time course of monocultures of BS4 (**a**) and BS5 (**b**) in white light. Competition between BS4 and BS5 in red light (**c**), green light (**d**) and white light (**e**). Symbols represent the observed population densities of BS4 (green triangles) and BS5 (red triangles). Lines represent the population densities predicted by the model: green solid line for BS4, red dashed line for BS5. Population densities are expressed in biovolumes (in femtolitres) per millilitre of water.

data (Fig. 2a, b). Parameter values thus obtained are given in Table 1. In a next step, we ran competition experiments with BS4 and BS5 under red light, green light and white light (full spectrum). The population dynamics in these competition experiments are compared against model simulations of competition using the parameter values estimated from the monocultures. Under red light, only BS4 (the green species) was able to survive (Fig. 2c). Under green light, BS5 (the red species) was the only survivor (Fig. 2d). In both cases, the experimentally observed population densities remained somewhat below the predicted population densities, presumably because photosynthesis is generally less efficient when all available light is concentrated on a single pigment¹⁸. Under white light, competition yielded stable coexistence of BS4 and BS5 for at least 60 days, in excellent agreement with the model predictions (Fig. 2e). Simulation of the model over a longer time-span predicted that BS4 and BS5 would maintain almost equal abundances in the long run. These findings demonstrate that partitioning of the light spectrum favours species coexistence.

Some species are able to adapt their pigment composition to the prevailing light spectrum. *Tolypothrix tenuis*, for example, is a marine filamentous cyanobacterium that is able to adjust the ratio of its phycocyanin (PC) to phycoerythrin (PE) content while keeping the total amount of these two pigments constant. This phenomenon is known as complementary chromatic adaptation^{24,25}. We incorporated this ability of chromatic adaptation in the model by the assumption that the ratio between phycoerythrin and phycocyanin changes in a direction that results in an increased specific growth rate under the prevailing light conditions (that is, an increased fitness)^{26,27}. That is, if x_T denotes the fraction PC/(PC + PE) in *Tolypothrix*, then the adaptive dynamics of the pigment composition in *Tolypothrix* can be described as

$$\frac{dx_T}{dt} = \alpha_T \frac{\Phi_T}{z_m} \int_0^{z_m} \frac{\partial \gamma_T(x_T, z)}{\partial x_T} dz \quad (4)$$

where the constant of proportionality α_T is a measure of the rate of chromatic adaptation of *Tolypothrix*. To test these adaptive dynamics, we ran competition experiments with *Tolypothrix* against BS4 and BS5 under white light. *Tolypothrix* and BS4 were able to coexist (Fig. 3a), because *Tolypothrix* produced the complementary pigment phycoerythrin (Fig. 3b). That is, *Tolypothrix* turned red in the presence of the green cyanobacterium BS4. In competition with BS5, *Tolypothrix* was driven to low abundance but BS5 was not able to exclude *Tolypothrix* (Fig. 3c). Hence, *Tolypothrix* and BS5 co-

Table 1 Parameter values and their interpretation

Symbol	Interpretation	Units	Value
Independent variables			
t	Time	h	–
z	Depth	cm	–
λ	Wavelength	nm	–
Dependent variables			
N_i	Population density of species i	fl cm ⁻³	–
$\gamma_i(z)$	Absorbed photons by species i	μmol photons s ⁻¹ fl ⁻¹	–
$I(\lambda, z)$	Underwater light spectrum	μmol photons m ⁻² s ⁻¹ nm ⁻¹	–
x_T	Fraction of phycocyanin in <i>Tolypothrix</i>	Dimensionless	–
Parameters			
$I_{in}(\lambda)$	Spectrum of incident light	μmol photons m ⁻² s ⁻¹ nm ⁻¹	m
$\int_{400}^{700} I_{in}(\lambda) d\lambda$	PAR-integrated incident light intensity	μmol photons m ⁻² s ⁻¹	40 m
$K_{bg}(\lambda)$	Spectrum background absorption	cm ⁻¹	m
$k_i(\lambda)$	Absorption spectrum of species i	cm ² fl ⁻¹	m
z_m	Depth of the water column	cm	7.7 m
L_i	Specific loss rate of species i	h ⁻¹	0.014 m
ϕ_i	Photosynthetic efficiency of:	fl (μmol photons) ⁻¹	
	BS4		2.2 × 10 ⁶ e
	BS5		1.6 × 10 ⁶ e
	<i>Tolypothrix</i>		1.7 × 10 ⁶ e
α_T	Chromatic adaptation parameter of <i>Tolypothrix</i>	Dimensionless	0.12 e

fl, femtolitre, denoting the biovolume of the species. m, measured parameter; e, estimated parameter (see Methods).

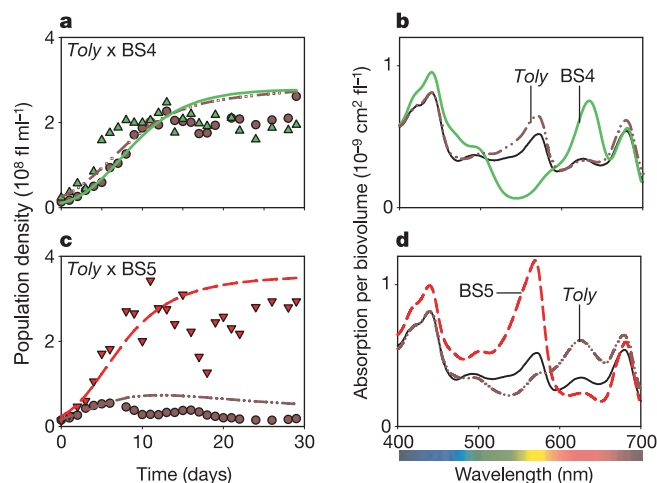


Figure 3 Competition between the filamentous cyanobacterium *Tolypothrix* and the picocyanobacteria BS4 and BS5. **a**, Competition between *Tolypothrix* and BS4. **b**, Absorption spectra of BS4 (green solid line), and of *Tolypothrix* at day 0 (black solid line) and at day 30 (brown dash-dotted line) of the competition experiment. **c**, Competition between *Tolypothrix* and BS5. **d**, Absorption spectra of BS5 (red dashed line), and of *Tolypothrix* at day 0 (black solid line) and at day 30 (brown dash-dotted line) of the competition experiment. In **a** and **c**, symbols represent the observed population densities of BS4 (green triangles), BS5 (red triangles), and *Tolypothrix* (filled brown circles); lines represent the population densities predicted by the model: green solid line for BS4, red dashed line for BS5, and brown dash-dotted line for *Tolypothrix*. Population densities are expressed in biovolumes (in femtolitres) per millilitre of water.

existed as well, because in this case *Tolypothrix* produced the complementary pigment phycocyanin (Fig. 3d). That is, *Tolypothrix* turned green in the presence of the red cyanobacterium BS5. The model predictions agreed well with the experimental results (Figs 3a,c). Furthermore, simulation of the model over a longer timespan predicted that stable coexistence would be maintained in the long run. These findings show that complementary chromatic adaptation favours divergence in pigment composition of competing species.

The diversity of the phytoplankton is commonly attributed to a variety of factors, including differential nutrient utilization, predation resistance, spatial and temporal heterogeneity, and complex dynamics^{1–5}. Recent discoveries of various new groups of phototrophic microorganisms in the oceans^{6–13} have spurred the novel hypothesis that differences in photosynthetic characteristics may offer subtle opportunities for niche differentiation of the plankton as well^{14–17}. Our findings provide firm evidence for this hypothesis. The theory and experiments demonstrate that selective forces promote partitioning of the light spectrum, which favours divergence in pigment composition. This divergence allows a more efficient utilization of the available light energy, and contributes to the unexpected biodiversity of phototrophic microorganisms in aquatic ecosystems. □

Methods

Experiments

Experiments were performed in chemostats specifically designed to study light-limited growth²¹. Fluorescent tubes with a white spectrum (Philips TLD 18W/965) were used as light source in all experiments. Green light and red light was obtained by placing coloured filters between the light source and the chemostat vessels (Lee no. 124 dark green filter for green conditions, and Lee no. 26 red filter for red conditions). In all experiments, the light intensity incident upon the chemostat vessels was set at $40 \mu\text{mol photons m}^{-2} \text{s}^{-1}$ (integrated over PAR range), as measured with a Licor LI-190SA quantum sensor. The spectrum of the incident light, $I_{\text{in}}(\lambda)$, was measured by a Licor LI-1800 spectroradiometer. To resemble the salinity of the Baltic Sea, we used the following brackish mineral medium: NaCl (8.25 g l^{-1}), $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$ (0.66 g l^{-1}), KCl (0.17 g l^{-1}), $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ (1.16 g l^{-1}), $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$ (0.17 g l^{-1}), $\text{Na}_3\text{-citrate}$ (4.98 mg l^{-1}), $\text{Na}_2\text{-EDTA}$ (0.83 mg l^{-1}), NaNO_3 (1.25 g l^{-1}), trace metal mix (1.0 mg l^{-1}), Na_2CO_3 (46.0 mg l^{-1}), $\text{K}_2\text{HPO}_4 \cdot 3\text{H}_2\text{O}$

(33.2 mg l^{-1}), $\text{Fe-NH}_4\text{-citrate}$ (4.8 mg l^{-1}). Given the maximal specific growth rates of BS4, BS5 and *Tolypothrix* in the range $0.025\text{--}0.030 \text{ h}^{-1}$, the chemostats were run at a dilution rate of 0.014 h^{-1} .

Samples were taken almost every day and population densities were determined. The spherical cells of BS4 and BS5 were counted with a Coulter Elite flow cytometer, which discriminated between the two species on the basis of their pigment fluorescence²⁸. The filamentous cyanobacterium *Tolypothrix* was counted by microscope using image analysis software (Leica Qwin standard, Version 2.5). The absorption spectra of the species, $k_i(\lambda)$, were measured by an Aminco DW-2000 double-beam spectrophotometer. To monitor the changes in the absorption spectra of *Tolypothrix* during the competition experiments with BS4 and BS5, *Tolypothrix* was isolated from the species mixture by percoll-gradient centrifugation (Amersham Biosciences). The absorption spectrum of the background, $K_{\text{bg}}(\lambda)$, is calculated on the basis of the spectrum of the incident light and the spectrum of light that leaves the chemostat vessels when filled with mineral medium only.

Simulations

Numerical simulations are based on a fourth order Runge-Kutta procedure for time integration, and Simpson's rule for depth integration. Parameter values are given in Table 1. The spectra of incident light and background absorption, the absorption spectra of the species, and the PAR-integrated incident light intensity were measured as described above. The depth of the water column equalled the depth of the chemostat vessel. The specific loss rate of the species was set equal to the dilution rate of the chemostat. The photosynthetic efficiencies of the species were estimated by fitting the model to monoculture experiments. The chromatic adaptation parameter of *Tolypothrix* was estimated by fitting the model to experiments with *Tolypothrix* under a changing light spectrum. These model fits were obtained by minimization of the residual sum of squares by means of the Gauss-Marquardt-Levenberg algorithm. This was performed by a software package called PEST (Watermark Numerical Computing). The parameters thus estimated were used to predict the population dynamics in the competition experiments.

Received 15 July; accepted 20 September 2004; doi:10.1038/nature03044.

Published online 10 October 2004.

- Hutchinson, G. E. The paradox of the plankton. *Am. Nat.* **95**, 137–145 (1961).
- Tilman, D. *Resource Competition and Community Structure* (Princeton Univ. Press, Princeton, 1982).
- Sommer, U. Comparison between steady state and non-steady state competition: experiments with natural phytoplankton. *Limnol. Oceanogr.* **30**, 335–346 (1985).
- Huisman, J. & Weissing, F. J. Biodiversity of plankton by species oscillations and chaos. *Nature* **402**, 407–410 (1999).
- Irigoin, X., Huisman, J. & Harris, R. P. Global biodiversity patterns of marine phytoplankton and zooplankton. *Nature* **429**, 863–867 (2004).
- Waterbury, J. B., Watson, S., Guillard, R. R. L. & Brand, L. E. Widespread occurrence of a unicellular, marine, planktonic cyanobacterium. *Nature* **277**, 293–294 (1979).
- Chisholm, S. W. *et al.* A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334**, 340–343 (1988).
- Moon-Van der Staay, S. Y., De Wachter, R. & Vaulot, D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001).
- Kolber, Z. S., Van Dover, C. L., Niederman, R. A. & Falkowski, P. G. Bacterial photosynthesis in surface waters of the open ocean. *Nature* **407**, 177–179 (2000).
- Kolber, Z. S. *et al.* Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean. *Science* **292**, 2492–2495 (2001).
- Béjà, O. *et al.* Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**, 630–633 (2002).
- Béjà, O. *et al.* Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906 (2000).
- Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
- Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**, 464–467 (1998).
- Béjà, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. Proteorhodopsin phototrophy in the ocean. *Nature* **411**, 786–789 (2001).
- Rocap, G. *et al.* Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042–1047 (2003).
- Ernst, A., Becker, S., Wollenzien, U. I. A. & Postius, C. Ecosystem-dependent adaptive radiations of picocyanobacteria inferred from 16S rRNA and ITS-1 sequence analysis. *Microbiology* **149**, 217–228 (2003).
- Kirk, J. T. O. *Light and Photosynthesis in Aquatic Ecosystems* 2nd edn (Cambridge Univ. Press, Cambridge, 1994).
- Falkowski, P. G. & Raven, J. A. *Aquatic Photosynthesis* (Blackwell Science, Malden, MA, 1997).
- Huisman, J. & Weissing, F. J. Light-limited growth and competition for light in well-mixed aquatic environments: an elementary model. *Ecology* **75**, 507–520 (1994).
- Huisman, J., Jonker, R. R., Zonneveld, C. & Weissing, F. J. Competition for light between phytoplankton species: experimental tests of mechanistic theory. *Ecology* **80**, 211–222 (1999).
- Huisman, J. *et al.* Changes in turbulent mixing shift competition for light between phytoplankton species. *Ecology* **85**, 2960–2970 (2004).
- Sathyendranath, S. & Platt, T. Computation of aquatic primary production: extended formalism to include effect of angular and spectral distribution of light. *Limnol. Oceanogr.* **34**, 188–198 (1989).
- Tandeau de Marsac, N. Occurrence and nature of chromatic adaptation in cyanobacteria. *J. Bact.* **130**, 82–91 (1977).
- Ohki, K., Gannt, E., Lippschultz, C. A. & Ernst, M. C. Constant phycobilisome size in chromatically adapted cells of the cyanobacterium *Tolypothrix tenuis*, and variation in *Nostoc* sp. *Plant Physiol.* **79**, 943–948 (1985).
- Metz, J. A. J., Nisbet, R. M. & Geritz, S. A. H. How should we define fitness for general ecological scenarios? *Trends Ecol. Evol.* **7**, 198–202 (1992).
- Abrams, P. A. The adaptive dynamics of consumer choice. *Am. Nat.* **153**, 83–97 (1999).

28. Jonker, R. R., Meulemans, J. T., Dubelaar, G. B. J., Wilkins, M. F. & Ringelberg, J. Flow cytometry: a powerful tool in analysis of biomass distributions in phytoplankton. *Water Sci. Technol.* **32**, 177–182 (1995).

Acknowledgements We thank M. Staal for his contribution to the isolation of the Baltic Sea picocyanobacteria, C. Sigon for support during the competition experiments, B. Sommeijer for advice on efficient numerical techniques, and O. Béja for comments on the manuscript. The research of M.S. and J.H. was supported by the Earth and Life Sciences Foundation (ALW), which is subsidized by the Netherlands Organisation for Scientific Research (NWO).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.H. (jef.huisman@science.uva.nl).

Baf60c is essential for function of BAF chromatin remodelling complexes in heart development

Heiko Lickert^{1,*}, Jun K. Takeuchi^{2,3,*}, Ingo von Both¹, Johnathon R. Walls^{4,5}, Fionnuala McAuliffe^{1,6}, S. Lee Adamson^{1,3,7}, R. Mark Henkelman^{4,5}, Jeffrey L. Wrana^{1,8}, Janet Rossant^{1,8} & Benoît G. Bruneau^{2,3,8}

¹Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada

²Cardiovascular Research and Developmental Biology, ⁴Mouse Imaging Centre, The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

³Heart and Stroke/Richard Lewar Centre of Excellence, ⁵Department of Medical Biophysics, ⁸Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada

⁶University College Dublin, Department of Obstetrics & Gynaecology, National Maternity Hospital, Holles Street, Dublin 2, Ireland

⁷Department of Obstetrics and Gynecology, University of Toronto, Toronto, Ontario M5G 1L4, Canada

* These authors contributed equally to this work

Tissue-specific transcription factors regulate several important aspects of embryonic development. They must function in the context of DNA assembled into the higher-order structure of chromatin. Enzymatic complexes such as the Swi/Snf-like BAF complexes remodel chromatin to allow the transcriptional machinery access to gene regulatory elements^{1,2}. Here we show that *Smarcd3*, encoding Baf60c, a subunit of the BAF complexes, is expressed specifically in the heart and somites in the early mouse embryo. *Smarcd3* silencing by RNA interference in mouse embryos derived from embryonic stem cells causes defects in heart morphogenesis that reflect impaired expansion of the anterior/secondary heart field, and also results in abnormal cardiac and skeletal muscle differentiation. An intermediate reduction in *Smarcd3* expression leads to defects in outflow tract remodelling reminiscent of human congenital heart defects. Baf60c overexpressed in cell culture can mediate interactions between cardiac transcription factors and the BAF complex ATPase Brg1, thereby potentiating the activation of target genes. These results reveal tissue-specific and dose-dependent roles for Baf60c in recruiting BAF chromatin remodelling complexes to heart-specific enhancers, providing a novel mechanism to ensure transcriptional regulation during organogenesis.

In a screen for targets of Wnt signalling (H.L., B. Cox, I.v.B., L. Attisano, J.L.W., M. M. Taketo, R. Kemler, J.R., unpublished work) we identified *Smarcd3*, which encodes Baf60c, a 60-kDa

subunit of the BAF complexes³. *Smarcd3* mRNA and Baf60c protein are initially restricted to the developing heart, as early as embryonic day 7.5 (E7.5) (Fig. 1a–g). In the looping heart tube (E8.5–E9.0), *Smarcd3* expression is more pronounced at the poles of the heart, which will give rise to the outflow tract anteriorly and to the atria posteriorly (Supplementary Fig. 1). *Smarcd3* is also expressed

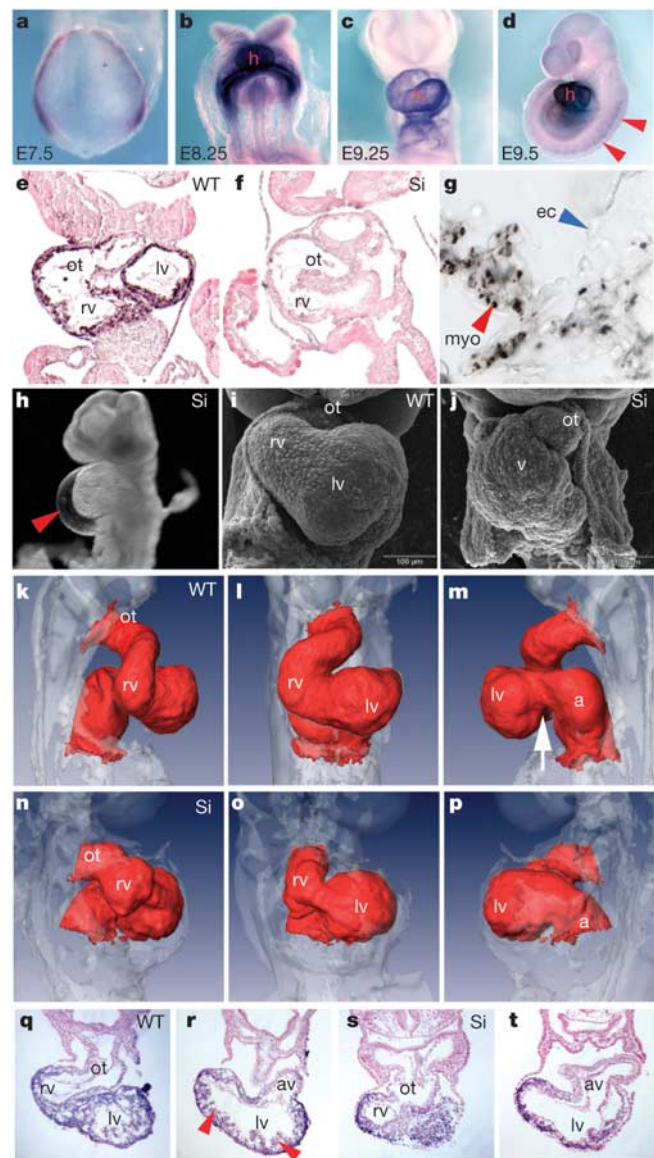
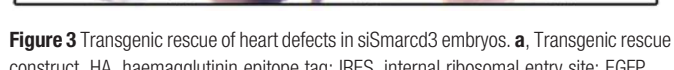


Figure 1 Baf60c is required for heart formation. **a–d**, *Smarcd3* is restricted to heart (h) at E7.75 (**a**), E8.25 (**b**) and E9.25 (**c**), and somites at E9.5 (arrowheads in **d**). **e–g**, Baf60c immunohistochemistry (brown) in E9.5 wild-type (WT; **e**, **g**) and *Smarcd3* knockdown (Si; **f**) embryos. Magnifications: $\times 10$ (**e**, **f**), $\times 40$ (**g**). **g**, Baf60c immunostaining in nuclei of myocardial cells (myo) but not endocardial cells (ec). **h**, Pericardial oedema in E8.75 *Smarcd3* knockdown embryo (arrowhead). **i**, **j**, Scanning electron micrographs of WT (**i**) and *Smarcd3* knockdown (Si; **j**) embryos at E8.75; a single ventricle (v) is apparent in the Si embryo. WT embryos have left ventricular (lv) and right ventricular (rv) chambers. **k–p**, Rendered OPT of E9.5 embryos viewed from the right (**k**, **n**), front (**l**, **o**), and left (**m**, **p**). The heart is red and the rest of the embryo is translucent. **k–m**, WT embryo. **n–p**, *Smarcd3* knockdown embryo, showing shortened outflow tract (ot), hypoplastic right ventricle (rv) and atrium (a), lack of atrio-ventricular canal (compare with **m**, arrow). **q–t**, Histology of E9.5 embryos stained by *in situ* hybridization with *Myl2* (blue). **q**, **r**, WT embryo; arrowheads show trabeculae. **s**, **t**, *Smarcd3* knockdown embryo. Note shortened outflow tract and lack of trabeculae (WT: arrowheads in **r**).

To address the role of *Smarcd3* *in vivo*, we employed embryonic stem (ES)-cell-mediated transgenic RNA interference (RNAi)⁵. Three *Smarcd3* ‘knockdown’ ES cell clones with less than 15%



108 ©2004 Nature



Publishing Group NATURE | VOL 432 | 4 NOVEMBER 2004 | www.nature.com/nature

To understand the basis for the cardiac defects in *Smarcd3* knockdown embryos we analysed markers of primary and anterior/secondary heart fields at the head-fold stage^{6,7}. The onset of expression of cardiac transcription factor *Nkx2-5*, marking both fields, was delayed (Fig. 2a, b; Supplementary Fig. 4). However, primary and secondary heart fields were initiated normally, as shown by the expression of *Gata4*, *Isl1*, *Fgf8* and *Fgf10* (Supplementary Fig. 4, and data not shown). During chamber morphogenesis (E8.5–E9.25), overall patterning and differentiation of the heart were unaffected, as the patterned expression of *Myl2* and *Tbx5* and the expression of *Actc* and *Mybpc3* were intact (Figs 1s, t and 2g, h; Supplementary Fig. 4). Expression of *Nppa*, marking the differentiating chamber myocardium, was reduced (Fig. 2c, d). The LV marker *Hand1* was expanded (Fig. 2e, f), whereas *Hand2*, a gene essential for formation of the RV⁸, was absent (Fig. 3d).

Trabecular formation was severely impaired in *Smarcd3* knockdown embryos (Fig. 1q–t). The trabeculae of the heart form as clonal outgrowths of the compact myocardium⁹. Expression of *Bmp10*, which is essential for trabecular differentiation¹⁰, and *Irx3*, which marks trabeculae, was undetectable in *Smarcd3* knockdown embryos (Fig. 2i, j; Supplementary Fig. 4). Myocardial thickness was increased threefold (40 μ m in *Smarcd3* knockdown embryos, compared with 12 μ m in wild-type embryos at E9.5), indicating that cells that normally differentiate into trabeculae remained part of the myocardial wall.

Consistent with severe outflow tract defects was our observation that the expression of genes involved in outflow tract formation (*Tbx20*, *Pitx2*, *Bmp4*, *Fgf8*, *Fgf10*, *Wnt11*) was impaired (Figs 2k–p and 3c; Supplementary Fig. 4; data not shown). TdT-mediated dUTP nick end labelling staining revealed no increased apoptosis

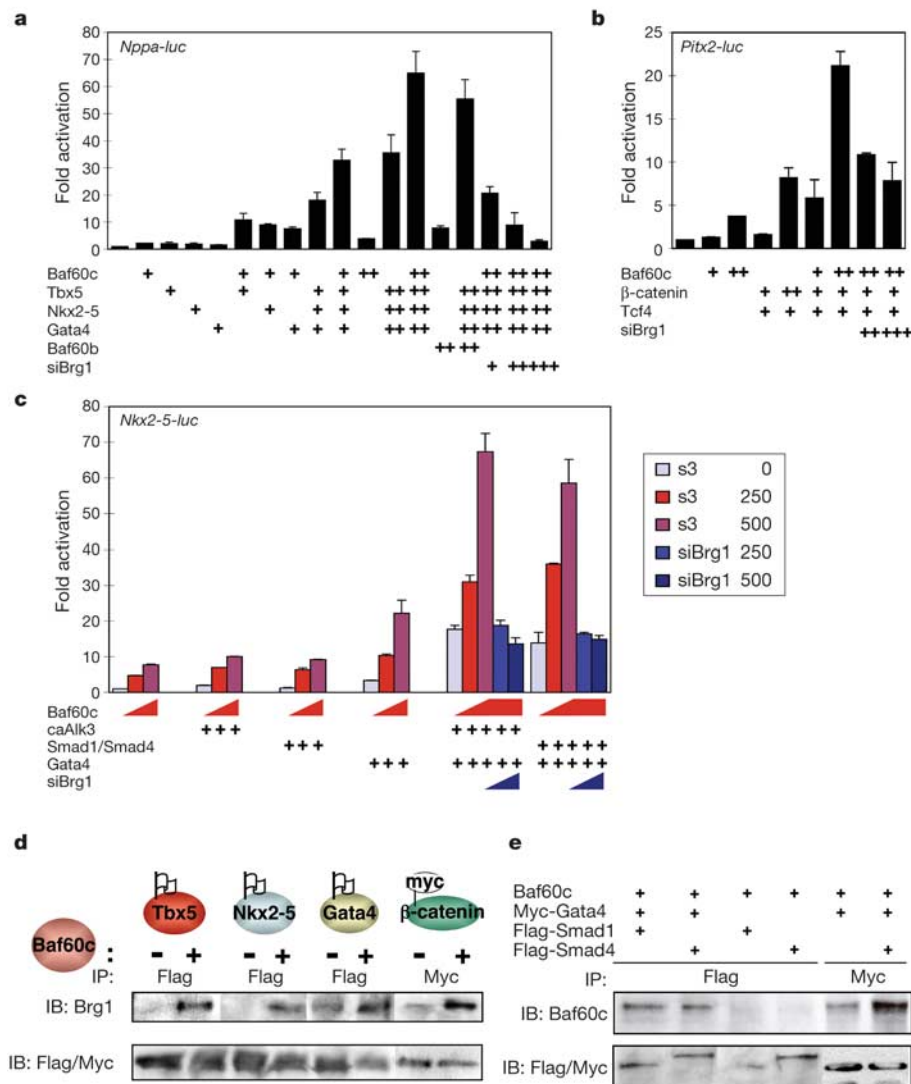


Figure 4 Baf60c potentiates transcription by promoting interactions between transcription factors and the BAF complex. **a**, Baf60c or Baf60b potentiates transactivation of *Nppa-luciferase* reporter construct by Tbx5, Nkx2-5 or Gata4, alone or in combination. **b**, Baf60c potentiates transactivation of *Pitx2-luciferase* reporter construct by activated β -catenin and Tcf4. **c**, Baf60c potentiates transactivation of *Nkx2-5-luciferase* reporter construct by Gata4 or Gata4 in combination with Smad1/Smad4 or caAlk3, but not activated BMP receptor (caAlk3) or Smad1/Smad4 alone. In **a–c**, one plus sign represents 50 ng of expression construct and two plus signs 100 ng, except for siBrg1. **a–c**, Increased transactivation by Baf60c is abrogated by siRNAs directed against *Brg1* mRNA (siBrg1; one plus sign, 100 ng; two plus signs, 250 ng;

three plus signs, 500 ng). **d**, Immunoprecipitation of transfected Flag-tagged Tbx5, Nkx2-5 or Gata4, or Myc-tagged β -catenin, followed by detection of Brg1 protein shows that Brg1 can only be detected in the presence of co-transfected Baf60c, or, for Gata4 and β -catenin, the interaction with Brg1 is strengthened by Baf60c. **e**, Co-immunoprecipitation of transfected Flag-tagged Smad1 or Smad4 or Myc-tagged Gata4, co-transfected with a Baf60c expression construct, followed by detection of Baf60c protein shows that Baf60c can interact with Gata4 but not with Smad1 or Smad4. Data are presented as means and s.e.m. (error bars) for three or four independent experiments.

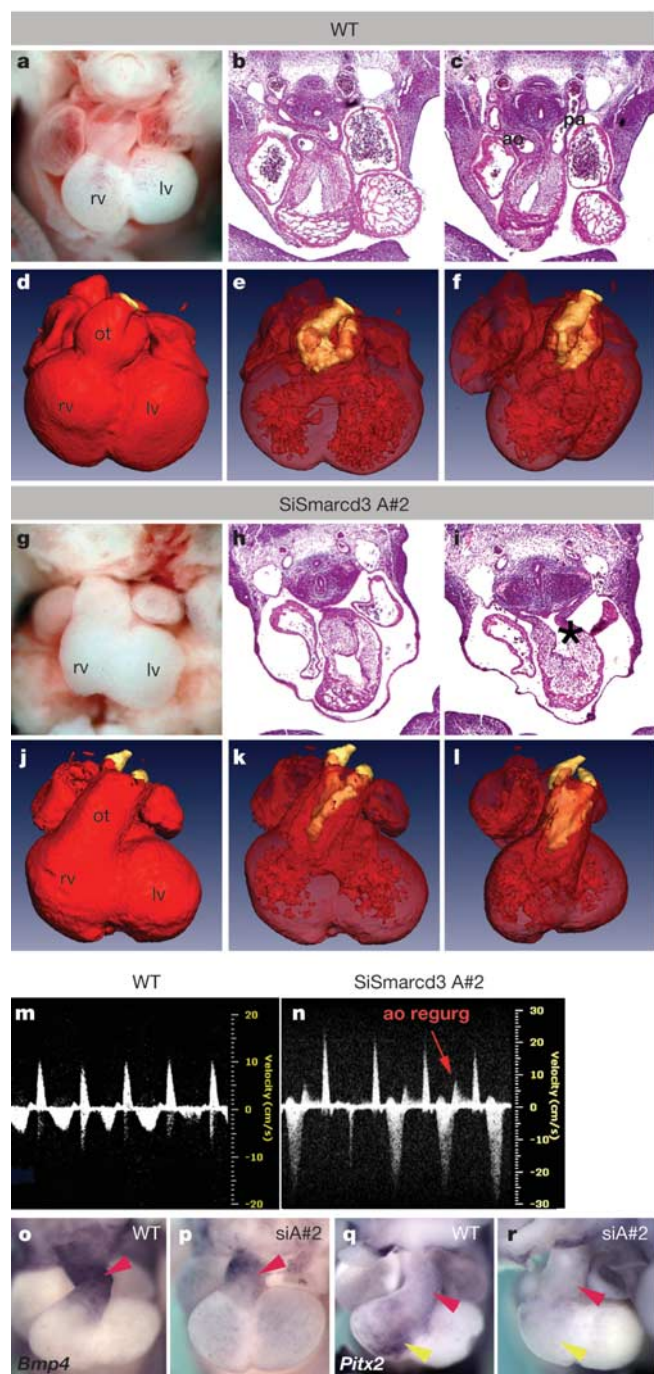


Figure 5 Dosage-sensitive role of Baf60c in outflow tract septation. **a–f**, Wild-type (WT) embryos at E12.5 (**a**, **d–f**) or E11.5 (**b**, **c**). **g–l**, *Smarcd3* intermediate knockdown (siSmarcd3 A#2) embryos at E12.5 (**g**, **j–l**) or E11.5 (**h**, **i**). WT embryos have a spiralling outflow tract, whereas siSmarcd3 A#2 embryos have a straight, unspiralled outflow tract, as shown by a brightfield view of fixed embryos (**a**, **g**), histology (**b**, **c**, **h**, **i**) and rendered OPT imaging (**d–f**, **j–l**) with the use of MF20 immunofluorescence to label cardiac muscle (red) and manually segmented PECAM immunofluorescence to label outflow tract endothelium (yellow). The opacity of the MF20 signal was decreased on the software display to permit the observation of the PECAM stain in **e**, **f**, **k** and **l**. The asterisk in **i** shows an unspiralled common outflow tract. **l**, **n**, Embryonic ultrasound demonstrates aortic regurgitation (ao regurg) in a siSmarcd3 A#2 embryo. **o–r**, Expression of *Bmp4* (**o**, **p**) and *Pitx2* (**q**, **r**) in wild-type (WT; **o**, **q**) and siSmarcd3 A#2 (siA#2; **p**, **r**) embryos. ao, aorta; lv, left ventricle; pa, pulmonary artery; rv, right ventricle. Red arrows show outflow tract, yellow arrows show RV.

(data not shown), but proliferation was significantly reduced in E9.5 *Smarcd3* knockdown myocardium as assessed by immunostaining for Ki67 (Fig. 2q, r). Because outflow tract, RV and atrium arise from the anterior/secondary heart field^{6,11}, these results suggest that failure of expansion of the anterior/secondary heart field underlies the major cardiac defects in *Smarcd3* knockdown embryos.

Skeletal muscle differentiation relies on an interacting cascade of transcriptional activators¹². Inhibiting the function of Brg1, the major catalytic subunit of the BAF complexes, prevents the myogenic actions of MyoD, Myf5 and MRF4 in cultured cells, and thus BAF complexes are thought to act in muscle fate determination by promoting the actions of myogenic transcription factors^{13,14}. Muscle differentiation was impaired in the absence of *Smarcd3* (Supplementary Fig. 5). Although the expression of upstream myogenic genes such as *Myf5* was intact, downstream genes were not expressed, including *Mef2c*, *Myogenin* and *Actc* (Supplementary Fig. 5 and data not shown). *Actc* expression was intact in the heart, suggesting that Baf60c acts on the *Actc* muscle, but not heart, enhancer¹⁵. These *in vivo* results directly corroborate the proposed role for BAF complexes in promoting muscle differentiation, and show that Baf60c is a crucial mediator.

To test whether members of the Baf60 family are functionally interchangeable, we performed a transgenic rescue experiment. We used the cytomegalovirus (CMV) enhancer/chicken β -actin promoter (pCAGG), which often drives cardiac-restricted expression¹⁶, to express Baf60b in the heart (Fig. 3a). In siSmarcd3-A#1 embryos in which cardiac-specific expression of Baf60b was achieved, heart formation and gene expression were rescued, whereas somite and neural tube defects remained (Fig. 3b–d). This demonstrates that Baf60c functions can be conferred by Baf60b, and that the RNAi-mediated phenotypes are not due to off-target effects.

Nppa, *Pitx2* and *Nkx2-5* expression was reduced in *Smarcd3* knockdown embryos. We examined the role of Baf60c in their transcriptional activation by using transient reporter assays in 10T1/2 cells. The *Nppa* promoter was activated by combinatorial interactions of Tbx5, *Nkx2-5* and Gata4 (Fig. 4a)^{17–19}, and transfection of Baf60c or Baf60b with Tbx5, *Nkx2-5* or Gata4 expression constructs individually or in combination led to synergistic activation of *Nppa-luciferase* (Fig. 4a). *Pitx2* expression in the outflow tract is regulated by Wnt signalling through Lef/Tcf binding sites²⁰; activation of *Pitx2-luciferase* by activated β -catenin and Tcf4 was also potentiated by Baf60c (Fig. 4b). *Nkx2-5* expression in the cardiac crescent relies on a Smad and GATA-factor-binding element²¹, and *Nkx2-5-luciferase* is strongly activated by Gata4, in synergy with bone morphogenetic protein (BMP) signalling and Smad1/Smad4 (ref. 21 and Fig. 4c). Baf60c potentiated the Gata4-dependent activation of *Nkx2-5-luciferase*, but had minimal impact on BMP signalling and Smad1/Smad4-dependent effects, indicating that Baf60c functions by means of interactions with Gata4 but not with Smads. Inhibition of Brg1 by RNAi eliminated the synergistic activation conferred by Baf60c, showing that Baf60c depends on the presence of Brg1 (Fig. 4a–c).

To determine whether the Brg1-dependent activation conferred by Baf60c relied on direct interactions between BAF complexes and cardiac transcription factors, we transfected HeLa cells with epitope-tagged β -catenin, Tbx5, *Nkx2-5* or Gata4 expression constructs, in the absence or presence of a Baf60c expression construct. Immunoprecipitation with antibodies directed against the epitope tags followed by immunoblotting for Brg1 revealed weak interactions of Brg1 with Gata4 and β -catenin as expected^{22,23}, and no detectable interaction with Tbx5 or *Nkx2-5* (Fig. 4d). However, expression of Baf60c induced an interaction between Brg1 and both Tbx5 and *Nkx2-5* and enhanced the association with Gata4 and β -catenin (Fig. 4d). Baf60c can therefore enhance interactions between transcription factors and Brg1, at least when overexpressed in cell culture. Next we examined interactions with Smads by co-immunoprecipitation and observed that Baf60c interacted with

Gata4, but not Smad1 nor Smad4 (Fig. 4e), which was consistent with our transactivation results. Because Smads recruit histone acetyltransferases (HATs) to activate transcription, we suggest a multi-step model in which BAF complexes, through interactions with Smad partners such as Gata4, synergize with HATs to allow the transition of chromatin into a transcriptionally active structure in response to transforming growth factor- β signalling²⁴. We conclude that Baf60c can promote interactions between specific cardiac transcription factors and BAF complexes, thereby potentiating the activation of heart-specific genes.

Our results show that Baf60c operates during cardiac development within pathways that regulate expansion of the heart at its poles^{6,7}. Baf60c interaction with Gata4 on the *Nkx2-5* cardiac crescent enhancer explains the similar cardiac defects in *Smarcd3* knockdown and *Nkx2-5*-null embryos (outflow tract defects, single ventricle and impaired trabeculation)^{25,26}. An affinity of Baf60c for multiple transcription factors or specific genomic loci might provide further degrees of specificity, explaining the complex cardiac phenotype of *Smarcd3* knockdown embryos.

We used the opportunity provided by transgenic RNAi to generate an epiallelic series^{5,27}. For this we used the siSmarcd3-A#2 line, which had 20% wild-type *Smarcd3* mRNA amounts in ES-cell-derived cardiomyocytes and gave rise to embryos with about 50% Baf60c protein amounts (Supplementary Fig. 3). These embryos survived until E13.0, and although overall heart formation seemed normal at E11.5, mutant hearts exhibited pericardial oedema, indicating heart failure (Fig. 5b, c, h, i). At E11.5 and E12.5 the spiralling septation of the outflow tract cushions did not occur, resulting in a straight and partly septated outflow tract (Fig. 5a–l), similar to human congenital malformations such as incomplete truncus arteriosus. We examined cardiac function *in utero* at E11.5 and found that 5 of 11 siSmarcd3-A#2 hearts had prominent outflow tract regurgitation (compared with 0 of 24 wild-type embryos), which is consistent with the outflow tract anomalies (Fig. 5m, n); siSmarcd3-A#2 embryos had lower heart rates and poorer myocardial function (Supplementary Table 1). Expression of *Pitx2* and *Bmp4*, genes important for outflow tract remodelling^{20,28}, was partly reduced in siSmarcd3-A#2 hearts (Fig. 5o–r), demonstrating dose-dependent gene regulation by Baf60c in the outflow tract. The phenotypes associated with the epiallelic series of reduction in *Smarcd3* mRNA amounts raise the possibility that dominant inherited congenital heart defects in humans might be caused by mutations in *Smarcd3* or other genes encoding BAF complex subunits.

Our results reveal tissue-specific and dose-dependent roles for Baf60c in recruiting BAF chromatin remodelling complexes to gene regulatory elements. Because all Baf60 isoforms are expressed in a tissue-specific manner, interactions between transcription factors and chromatin remodelling complexes might be a general feature of higher-order chromatin activation and gene regulation during tissue morphogenesis. □

Methods

In vivo RNAi and transgenesis

In vivo RNA interference was performed essentially as described in ref. 5. R1 ES cells were stably transfected with a plasmid vector containing an H1 promoter driving the expression of short hairpin RNAs (shRNAs). A neomycin resistance cassette was used for selection. Two individual shRNA sequences (siSmarcd3-A and siSmarcd3-B) were designed to distinct regions of the *Smarcd3* mRNA (see Supplementary Fig. 3). BLAST searches ensured that they were specific to *Smarcd3*. For selection of ES cells with reduced *Smarcd3* mRNA amounts, ES cells were differentiated into an enriched population of cardiac myocytes⁷, and real-time reverse-transcriptase-mediated polymerase chain reaction quantified *Smarcd3* mRNA. Tetraploid aggregations were performed with knockdown ES cells to generate embryos composed entirely of transgenic ES cells⁵. Tetraploid cells expressed enhanced green fluorescent protein (EGFP), thereby ensuring that only embryos totally derived from ES cells were used in the study. Transgenic rescue was performed by transfecting line siSmarcd3-A#1 with an expression construct composed of the CMV enhancer/chicken β -actin promoter (pCAGG) driving a haemagglutinin (HA)-tagged Baf60b cDNA, followed by an internal ribosomal entry site and EGFP¹⁶; a puromycin

resistance cassette was used for selection. Stably transfected ES cells were selected, and transgene expression was assessed by anti-HA and anti-GFP western blotting. ES cell lines with various degrees of Baf60b expression were used in tetraploid aggregations as described above. Some lines with Baf60b expression showed predominantly cardiac expression, which is characteristic of the pCAGG enhancer/promoter¹⁶.

In situ hybridization and antibody staining

All embryos were stage-matched to controls by somite count. Whole-mount or section *in situ* hybridization was performed in accordance with standard protocols. For fluorescent whole-mount *in situ* hybridization, embryos were hybridized with a mixture of digoxigenin-labelled *Myhpc3* and *Actc* probes, and detection was performed with rhodamine tyramide amplification reagents (Perkin-Elmer). Whole-mount immunofluorescence was done on embryos fixed in 4% paraformaldehyde, with the use of MF20 antiserum directly conjugated to Alexa 594 (Molecular Probes), and unconjugated rat anti-PECAM monoclonal antibody (Pharmingen) that was revealed with an Alexa 488-conjugated secondary antibody (Molecular Probes). The MF20 antibody developed by D. A. Fischman was obtained from the Developmental Studies Hybridoma Bank developed under the auspices of the National Institute of Child Health and Human Development and maintained by The University of Iowa (Department of Biological Sciences, Iowa City, Iowa, USA).

Optical projection tomography

Optical projection tomography (OPT) was performed essentially as described²⁹, on embryos fluorescently labelled by *in situ* hybridization or immunofluorescence. Visualization and manipulation of OPT data was performed with Amira software, version 3.0 (TGS Inc.). Segmentation was performed either automatically or, when necessary, manually.

Assessment of cardiac function *in utero*

Embryos in pregnant siSmarcd3-A#2 implanted mice and ICR wild-type mice were examined under isoflurane anaesthesia. An ultrasound biomicroscope (Visualsonics Inc., Toronto) with a transducer frequency of 30 MHz was used to image embryonic cardiac structures as described previously³⁰. Time intervals were obtained by placing a pulsed Doppler sample volume in the common ventricle to detect inflow through the atrio-ventricular valve and outflow through the common outflow tract. At the embryonic stage of examination, no E-wave (passive filling of the ventricle during atrial relaxation) was visible. We measured isovolumetric contraction time (ICT) as the interval between the end of the A-wave (active filling of the ventricle during atrial relaxation) and the onset of the outflow wave, ejection time (ET) as the interval between the start and end of the outflow wave, isovolumetric relaxation time (IVRT) as the interval between the end of the outflow wave and the onset of the A-wave, and atrio-ventricular conduction time as the interval between the onset of the A-wave and the onset of the outflow wave. The TEI index, (ICT + IVRT)/ET, is an index of global cardiac myocardial function that is independent of heart rate and ventricular morphology.

Transactivation assays and immunoprecipitation

Transactivation assays and co-immunoprecipitation experiments were performed essentially as described in refs 17 and 18. The *Nppa-luc* and *Nkx2-5-luc* (FL construct) reporters were described previously^{17,21}, and the *Pitx2-luc* reporter consisted of 2.5 kb of the upstream region of the *Pitx2* translational start site (C. Meno and H. Hamada, personal communication). Anti-Brg1 (Upstate), anti-Flag (Sigma), anti-HA (Sigma), anti-Myc (Santa Cruz) and anti-GFP (Molecular Probes) antisera were obtained commercially; anti-Baf60c antiserum was provided by M.-B. Debril and J. Auwerx; anti-Baf53a, anti-Baf155 and anti-Brg1 antisera were provided by W. Wang.

Received 24 August; accepted 30 September 2004; doi:10.1038/nature03071.

- Narlikar, G. J., Fan, H. Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475–487 (2002).
- Martens, J. A. & Winston, F. Recent advances in understanding chromatin remodeling by Swi/Snf complexes. *Curr. Opin. Genet. Dev.* **13**, 136–142 (2003).
- Wang, W. *et al.* Diversity and specialization of mammalian SWI/SNF complexes. *Genes Dev.* **10**, 2117–2130 (1996).
- Debril, M. B. *et al.* Transcription factors and nuclear receptors interact with the SWI/SNF complex through the BAF60c subunit. *J. Biol. Chem.* **279**, 16677–16686 (2004).
- Kunath, T. *et al.* Transgenic RNA interference in ES cell-derived embryos recapitulates a genetic null phenotype. *Nature Biotechnol.* **21**, 559–561 (2003).
- Cai, C. L. *et al.* Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Dev. Cell* **5**, 877–889 (2003).
- von Both, I. *et al.* FoxH1 is essential for development of the anterior heart field. *Dev. Cell* **7**, 331–345 (2004).
- Srivastava, D. *et al.* Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. *Nature Genet.* **16**, 154–160 (1997).
- Meilhac, S. M. *et al.* A retrospective clonal analysis of the myocardium reveals two phases of clonal growth in the developing mouse heart. *Development* **130**, 3877–3889 (2003).
- Chen, H. *et al.* BMP10 is essential for maintaining cardiac growth during murine cardiogenesis. *Development* **131**, 2219–2231 (2004).
- Meilhac, S. M., Esner, M., Kelly, R. G., Nicolas, J. F. & Buckingham, M. E. The clonal origin of myocardial cells in different regions of the embryonic mouse heart. *Dev. Cell* **6**, 685–698 (2004).
- Buckingham, M. Skeletal muscle formation in vertebrates. *Curr. Opin. Genet. Dev.* **11**, 440–448 (2001).
- de la Serna, I. L., Carlson, K. A. & Imbalzano, A. N. Mammalian SWI/SNF complexes promote MyoD-mediated muscle differentiation. *Nature Genet.* **27**, 187–190 (2001).
- Roy, K., de la Serna, I. L. & Imbalzano, A. N. The myogenic basic helix-loop-helix family of

- transcription factors shows similar requirements for SWI/SNF chromatin remodeling enzymes during muscle differentiation in culture. *J. Biol. Chem.* **277**, 33818–33824 (2002).
15. Biben, C., Hadchouel, J., Tajbakhsh, S. & Buckingham, M. Developmental and tissue-specific regulation of the murine cardiac actin gene *in vivo* depends on distinct skeletal and cardiac muscle-specific enhancer elements in addition to the proximal promoter. *Dev. Biol.* **173**, 200–212 (1996).
 16. Toyoda, M. *et al.* jumonji downregulates cardiac cell proliferation by repressing cyclin D1 expression. *Dev. Cell* **5**, 85–97 (2003).
 17. Durocher, D., Charron, F., Warren, R., Schwartz, R. J. & Nemer, M. The cardiac transcription factors Nkx2-5 and GATA-4 are mutual cofactors. *EMBO J.* **16**, 5687–5696 (1997).
 18. Bruneau, B. G. *et al.* A murine model of Holt–Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* **106**, 709–721 (2001).
 19. Garg, V. *et al.* GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* **424**, 443–447 (2003).
 20. Kioussi, C. *et al.* Identification of a Wnt/Dvl/ β -catenin $\rightarrow$ Pitx2 pathway mediating cell-type-specific proliferation during development. *Cell* **111**, 673–685 (2002).
 21. Brown, C. O. *et al.* The cardiac determination factor, Nkx2-5, is activated by mutual cofactors GATA-4 and Smad1/4 via a novel upstream enhancer. *J. Biol. Chem.* **279**, 10659–10669 (2004).
 22. Kadam, S. & Emerson, B. M. Transcriptional specificity of human SWI/SNF BRG1 and BRM chromatin remodeling complexes. *Mol. Cell* **11**, 377–389 (2003).
 23. Barker, N. *et al.* The chromatin remodeling factor Brg-1 interacts with β -catenin to promote target gene activation. *EMBO J.* **20**, 4935–4943 (2001).
 24. Derynck, R. & Zhang, Y. E. Smad-dependent and Smad-independent pathways in TGF- β family signalling. *Nature* **425**, 577–584 (2003).
 25. Lyons, I. *et al.* Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5. *Genes Dev.* **9**, 1654–1666 (1995).
 26. Tanaka, M., Chen, Z., Bartunkova, M., Yamazaki, N. & Izumo, S. The cardiac homeobox gene Csx/Nkx2.5 lies genetically upstream of multiple genes essential for heart development. *Development* **126**, 1269–1280 (1999).
 27. Hemann, M. T. *et al.* An epigenetic series of p53 hypomorphs created by stable RNAi produces distinct tumor phenotypes *in vivo*. *Nature Genet.* **33**, 396–400 (2003).
 28. Liu, W. *et al.* Bmp4 signaling is required for outflow-tract septation and branchial-arch artery remodeling. *Proc. Natl Acad. Sci. USA* **101**, 4489–4494 (2004).
 29. Sharpe, J. *et al.* Optical projection tomography as a tool for 3D microscopy and gene expression studies. *Science* **296**, 541–545 (2002).
 30. Zhou, Y. Q. *et al.* Applications for multifrequency ultrasound biomicroscopy in mice from implantation to adulthood. *Physiol. Genomics* **10**, 113–126 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Sharpe for providing a prototype OPT system, S. McMaster for tetraploid aggregations, D. Holmyard for scanning electron microscopy, K. Harpal and K. Koshiba-Takeuchi for histology, H. Hamada and C. Meno for the unpublished *Pitx2* reporter, and A. Nagafuchi, M. Nemer, R. Schwartz, D. Srivastava and T. Takeuchi for reporter and expression constructs. This research was funded by the Canadian Institutes of Health Research (CIHR; to B.G.B., J.R. and J.L.W.), the Heart and Stroke Foundation of Ontario (to B.G.B.), the March of Dimes Birth Defects Foundation (to B.G.B.), NCIC (to J.L.W.), the Canada Foundation for Innovation (to R.M.H.), an Emmy–Noether fellowship of the Deutsche Forschungsgemeinschaft (to H.L.), a Human Frontiers Science Program long-term fellowship (to J.K.T.), an OGSST fellowship (to J.R.W.), and the Koeln Fortune Program, University of Cologne (to L.v.B.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.G.B. (bbruneau@sickkids.ca) or J.R. (rossant@mshri.on.ca).

A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay

Nadia Amrani, Robin Ganesan, Stephanie Kervestin, David A. Mangus, Shubendu Ghosh & Allan Jacobson

Department of Molecular Genetics and Microbiology, University of Massachusetts Medical School, Worcester, Massachusetts 01655-0122, USA

Nonsense-mediated messenger RNA decay (NMD) is triggered by premature translation termination^{1–3}, but the features distinguishing premature from normal termination are unknown. One model for NMD suggests that decay-inducing factors bound to mRNAs during early processing events are routinely removed

by elongating ribosomes but remain associated with mRNAs when termination is premature, triggering rapid turnover⁴. Recent experiments^{5–7} challenge this notion and suggest a model that posits that mRNA decay is activated by the intrinsically aberrant nature of premature termination^{8,9}. Here we use a primer extension inhibition (toeprinting) assay¹⁰ to delineate ribosome positioning and find that premature translation termination in yeast extracts is indeed aberrant. Ribosomes encountering premature UAA or UGA codons in the *CAN1* mRNA fail to release and, instead, migrate to upstream AUGs. This anomaly depends on prior nonsense codon recognition and is eliminated in extracts derived from cells lacking the principal NMD factor, Upf1p, or by flanking the nonsense codon with a normal 3'-untranslated region (UTR). Tethered poly(A)-binding protein (Pab1p), used as a mimic of a normal 3'-UTR, recruits the termination factor Sup35p (eRF3) and stabilizes nonsense-containing mRNAs. These findings indicate that efficient termination and mRNA stability are dependent on a properly configured 3'-UTR.

The yeast *can1-100* allele contains a premature UAA codon at position 47 of the *CAN1* coding region that effectively terminates translation and destabilizes the *CAN1* mRNA¹¹. We constructed a gene that fused the UAA-containing segment of *can1-100* to the firefly *LUC* coding region, and used this construct to generate synthetic mRNA (UAA RNA). We also created a variant with a weak terminator (CAA UGA CAA)¹² at codon 47 (UGA RNA) and two control RNAs with no early stop codon (Fusion and AAA RNAs; Fig. 1a). *In vivo*, the UAA and UGA RNAs were substrates for NMD. All of these mRNAs were incubated in translation reactions with wild-type extracts and subjected to toeprinting analyses, using sensitivity to the cap analogue ⁷mGpppG, an inhibitor of translation initiation, as a method of distinguishing genuine toeprints from background bands¹⁰. Toeprints corresponding to ribosomes stalled with a stop codon in their A sites were obtained with the UAA and UGA RNAs at the expected position 12–14 nucleotides (nt) downstream of the premature nonsense codons¹⁰ (Fig. 1b, lanes 1 and 3). The toeprints were dependent on translation because they were sensitive to cap analogue (lanes 2 and 4), and were dependent on the presence of a stop codon (which is absent from the Fusion RNA; lanes 5 and 6). Similar experiments failed to detect any toeprint signals from the normal UAA termination codons of either Fusion or *ADE2* RNAs, even after long periods of translation (Supplementary Fig. 1a) or incubation in extracts with the temperature-sensitive termination factor Sup45p/eRF1 (data not shown). These results indicate that ribosomes that terminate prematurely are released much less efficiently than those encountering normal terminators.

Addition of the elongation inhibitor cycloheximide to the translation reactions also failed to reveal toeprints from normal terminators (data not shown) but did allow the detection of additional toeprints^{13,14}. Cycloheximide-dependent initiator AUG toeprints on the UAA, UGA, Fusion and AAA RNAs were sensitive to cap analogue (Fig. 1c, top arrow, lanes 1–6; Fig. 1d, top) and reflect 80S ribosomes, centred on AUG codons, protecting 16–18 nt 3' to those codons. Other cycloheximide-dependent toeprints were present in close proximity to the locations of the early stop codons. These unexpected bands mapped to a position 6–7 nt downstream of the terminator U in both the UAA and UGA RNAs (Fig. 1c, lanes 1 and 3, middle arrow), as well as to a position 17 nt downstream of the terminator U in the UGA RNA (lane 3, bottom arrow). The appearance of these toeprints was dependent on concurrent mRNA translation (lanes 2 and 4), the presence of yeast extract (lanes 7–9), and recognition of a termination codon (lanes 5 and 6; Fig. 1d, bottom). The dependence of these toeprints on prior termination codon recognition was underscored by experiments showing that, after the addition of cycloheximide, the +6 toeprint of the UAA RNA accumulated after the disappearance of the +12 toeprint

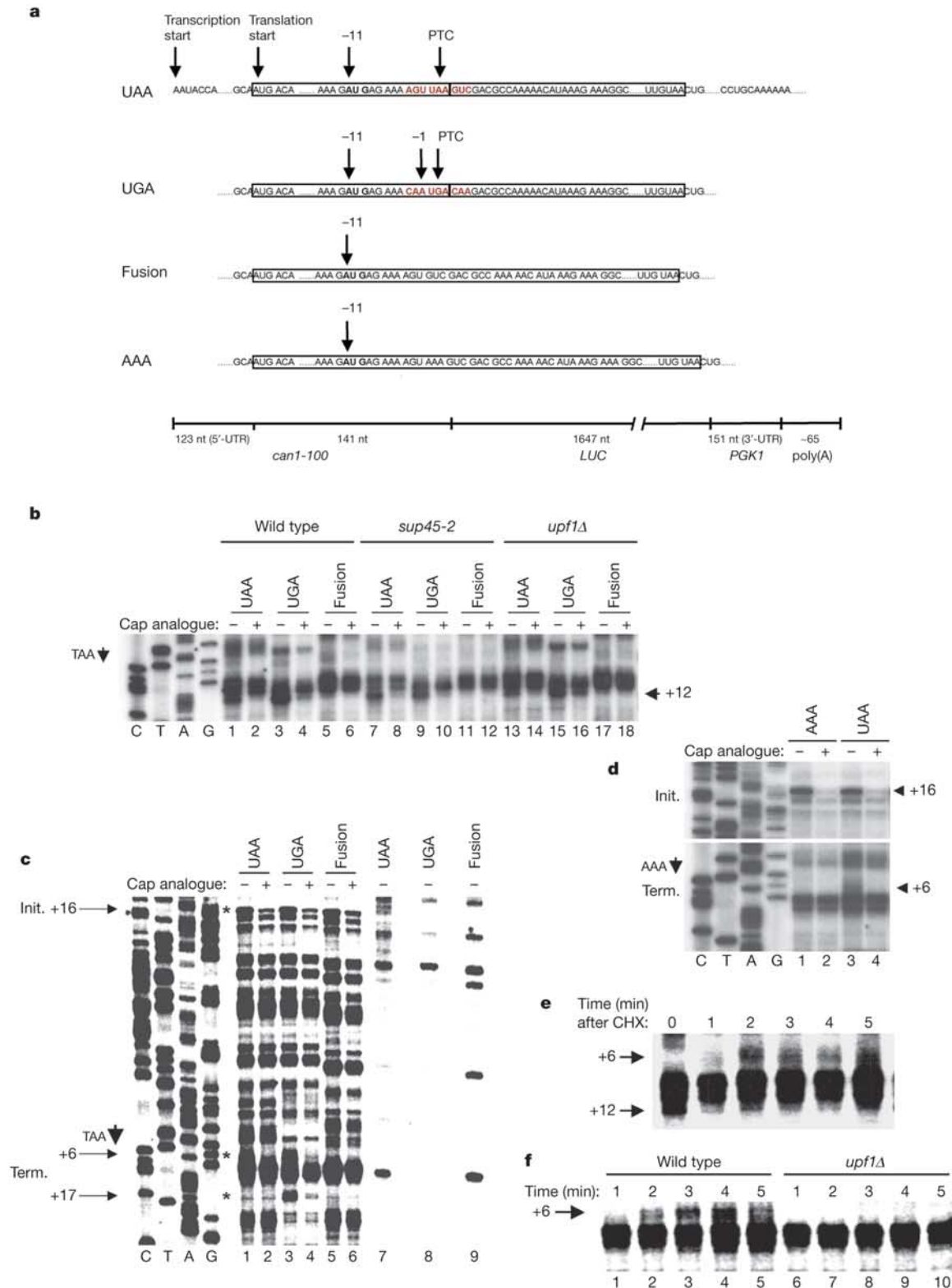


Figure 1 Toeprint analyses of initiation and premature termination in cell extracts. **a**, Selected regions of the UAA, UGA, Fusion and AAA *CAN1/LUC* RNAs. PTC, premature termination codon. PTCs and flanking sequences are indicated in red. **b**, PTC toeprints are detected during translation of *CAN1/LUC* RNAs in wild-type, *sup45-2* and *upf1Δ* extracts in the absence of cycloheximide. **c**, Addition of cycloheximide to wild-type extracts promotes the appearance of normal toeprints at initiator AUG codons and aberrant toeprints in the vicinity of PTCs. Lanes 7–9 show products from reactions lacking added extract. **d**, The +6 toeprint of the UAA RNA incubated in the presence of cycloheximide is

eliminated by a single nucleotide substitution. **e**, The +6 toeprint in the UAA RNA appears after cycloheximide-chasing of the PTC toeprint. Cycloheximide (CHX) was added after 4 min of translation, and samples were taken at the designated times. **f**, *upf1Δ* extracts do not reveal the +6 toeprint of the UAA RNA when incubated for 1–5 min and terminated by the addition of cycloheximide for 3 min. Positions of the toeprints are indicated with arrows. The left portions of **b–d** show dideoxynucleotide sequencing reactions for the UAA or AAA templates (with 5' → 3' sequence reading from top to bottom).

(Fig. 1e) and by the observation that extracts of *sup45-2* cells with defective eRF1 (ref. 15) yielded only the +12 toeprint, regardless of the presence or absence of cycloheximide (Fig. 1b, lanes 7–12; Supplementary Fig. 3a, lanes 7, 8, 17 and 18). Most importantly, the aberrant toeprints of the UAA and UGA RNAs were linked to NMD because they failed to accumulate in extracts lacking the NMD factors Upf1p or Nmd2p (Fig. 1f; data not shown). The lack of aberrant toeprints in *upf1Δ* or *nmd2Δ* extracts cannot be due to inefficient translation because these extracts display premature terminator +12 toeprints in the absence of cycloheximide (Fig. 1b, lanes 13–18; data not shown) and normal initiator AUG toeprints in the presence of cycloheximide (data not shown).

The UAA and UGA RNAs both contained AUG codons just upstream of their premature terminators (Fig. 1a). To determine whether these AUG codons were relevant to the toeprints of Fig. 1c–f, we analysed two RNAs in which the upstream AUG codons were mutated. Toeprint analyses of transcripts with mutated AUG codons (UAA-M and UGA-M RNAs) in wild-type extracts supplemented with cycloheximide, in the presence or absence of cap analogue, showed, first, that mutation of the –11 AUG codon of the UAA RNA eliminated the +6 band (Fig. 2, compare lanes 1 and 3) and, second, that mutation of the –11 and –1 AUG codons in the UGA RNA led to the disappearance of the +6 and +17 bands (compare lanes 7 and 9) and to the appearance of a +12 toeprint characteristic of weak terminators¹² (lane 9; data not shown). These results indicate that the +6 toeprints obtained with the UAA and UGA RNAs correspond to the P site of the ribosome stalled at the –11 AUG codon and that the +17 toeprint derived from the UGA RNA corresponds to ribosomes stalled at the –1 AUG codon (ref. 14). One explanation for these results is that post-termination ribosomes fail to be released at premature terminators and are able to scan backwards and reinitiate at the –11 AUG codon in UAA RNA or at the –11 or –1 AUG codon in UGA RNA. Such retro-reinitiation, previously suggested by studies in mammalian cells^{16,17}, was further supported by experiments showing that the luciferase activity obtained from an in-frame fusion of the –11 AUG codon to the *LUC* open reading frame (ORF) was fourfold to fivefold greater in the presence of the codon 47 UAA and the –11 AUG codon than in their absence (Supplementary Fig. 2). Although retro-reinitiation is alternatively explainable as leaky scanning¹⁴, the absence of a toeprint signal at the –11 AUG codon in Fusion or AAA RNAs or in *upf1Δ* or *nmd2Δ* extracts, as well as in the chase experiments of Fig. 1e, make this a less likely explanation. Collectively, these experiments indicate that access to internal AUG codons of UAA and UGA mRNAs depends on prior recognition of a downstream premature stop codon.

Construction of synthetic mRNAs with additional upstream AUGs showed that ribosomes could reinitiate *in vitro* at least 32 nt 5' to a premature terminator (Supplementary Fig. 1b). To test the relative efficiencies of upstream and downstream reinitiation, we constructed mRNAs that presented both options. Analysis of RNAs containing normal or mutated versions of both the –11 AUG codon and another AUG 5 nt downstream of the stop codon showed that ribosomes exiting from premature termination codons have a propensity for backwards scanning (Supplementary Fig. 1c). Regardless of the site of post-termination reinitiation, all nonsense-containing RNAs yielded toeprints corresponding to ribosomes stalled with the stop codon in their A sites when translated in *sup45-2* extracts (Supplementary Fig. 3a), indicating that, before any reinitiation event, a premature stop codon must be recognized by Sup45p/eRF1 and peptide hydrolysis must be triggered^{15,18}.

To determine whether aberrant termination and its consequences resulted from the relative positioning of a nonsense codon within an mRNA, four constructs were made in which the *PGK1* 3'-UTR^{19,20} was inserted immediately downstream of early stop codons in *can1* alleles. These 'mini' wild-type mRNA mimics included the miniUAA, miniUAA-M and miniUGA RNAs, as well the mini-

UAA-M-W RNA (which had the UAA in weak context¹² but lacked the upstream AUG codon) (Fig. 3a). All 'mini' RNAs yielded normal levels of initiator AUG toeprints in cycloheximide-supplemented wild-type extracts (data not shown) but lacked the aberrant +6 toeprint (Fig. 3b). Unlike the normal terminators of the Fusion and ADE2 RNAs, translation programmed by 'mini' RNAs in *sup45-2* extracts yielded cap-analogue-sensitive +14 toeprint signals in the absence of cycloheximide (Fig. 3c) but not in its presence (Supplementary Fig. 3b). The lack of aberrant toeprints at +6 and/or +17 with these mimics of wild-type mRNA indicates that backwards scanning, or at least its apparent end-product, is eliminated when termination codons are flanked by normal 3'-UTR sequences. These results indicate that aberrant termination promotes the novel initiation events and raise the question of why a 3'-UTR created by a premature termination codon should differ from that of a wild-type mRNA.

The 'faux UTR' model suggests that the downstream element (DSE) thought to be a key *cis*-acting regulator of NMD¹⁹ is a defective UTR that promotes mRNA decay because it lacks a termination regulatory factor (or factors) normally present on a legitimate 3'-UTR, a hypothesis that might also explain why deletions that eliminate most coding sequences downstream of premature terminators stabilize mRNAs that would otherwise be substrates for NMD^{19,20}. This inadequacy could occur because translation to the normal end of a coding region remodels a messenger ribonucleoprotein (mRNP)⁹ or because proximity to

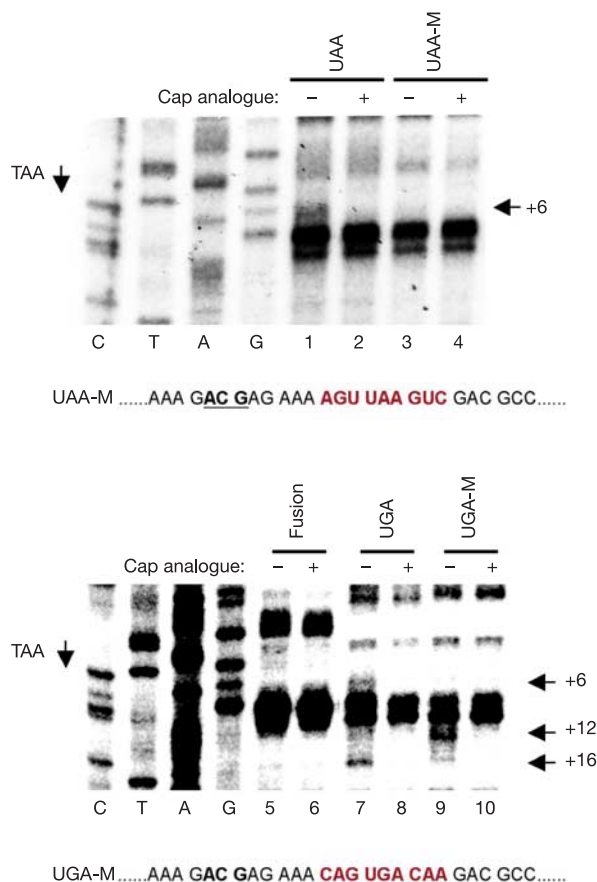


Figure 2 Aberrant toeprints derived from PTCs in wild-type extracts in the presence of cycloheximide are dependent on upstream AUGs. Toeprint analyses of the UAA, UAA-M, Fusion, UGA and UGA-M RNAs are as described in Fig. 1c. Pertinent sequences of mutated *CAN1/LUC* constructs are shown below each panel. PTCs and flanking sequences are indicated in red.

the poly(A) tail (and its bound Pab1p) has a qualitative and/or quantitative influence on the nature of proteins bound to the UTR⁸. To test whether proximity of a termination codon to Pab1p is germane to NMD, the *in vivo* stability of mRNAs bearing an MS2 coat protein-binding site 3' to premature terminators in the *CAN1* and *PGK1* mRNAs was assessed in cells expressing an MS2–Pab1p fusion. Pab1p tethered 37–73 nt 3' to premature UAA, UGA or UAG codons promoted 5–11-fold increases in mRNA stability and abundance (Supplementary Fig. 4a, b). As controls, MS2 dimer or MS2–Sxl (*sex lethal* protein) tethered at the same position, or MS2–Pab1p tethered 164 nt downstream of the stop codon and 3' to the DSE, had no effect on mRNA stability or abundance

(Supplementary Fig. 4b). Because the same number of ribosomes was associated with the mRNA in the presence or absence of tethered Pab1p (Supplementary Fig. 4c), the mRNA-stabilizing effects could not be attributed to selective interference with translation.

Tethered fragments of Pab1p caused partial stabilization of nonsense-containing transcripts (Supplementary Fig. 4d), suggesting that Pab1p-mediated mRNA stabilization might be attributable to protein–protein interactions characteristic of its respective domains²¹. To test this possibility, we used anti-MS2 coat protein antibodies to immunoprecipitate mRNPs and assayed for the presence of co-immunoprecipitated regulatory factors. In

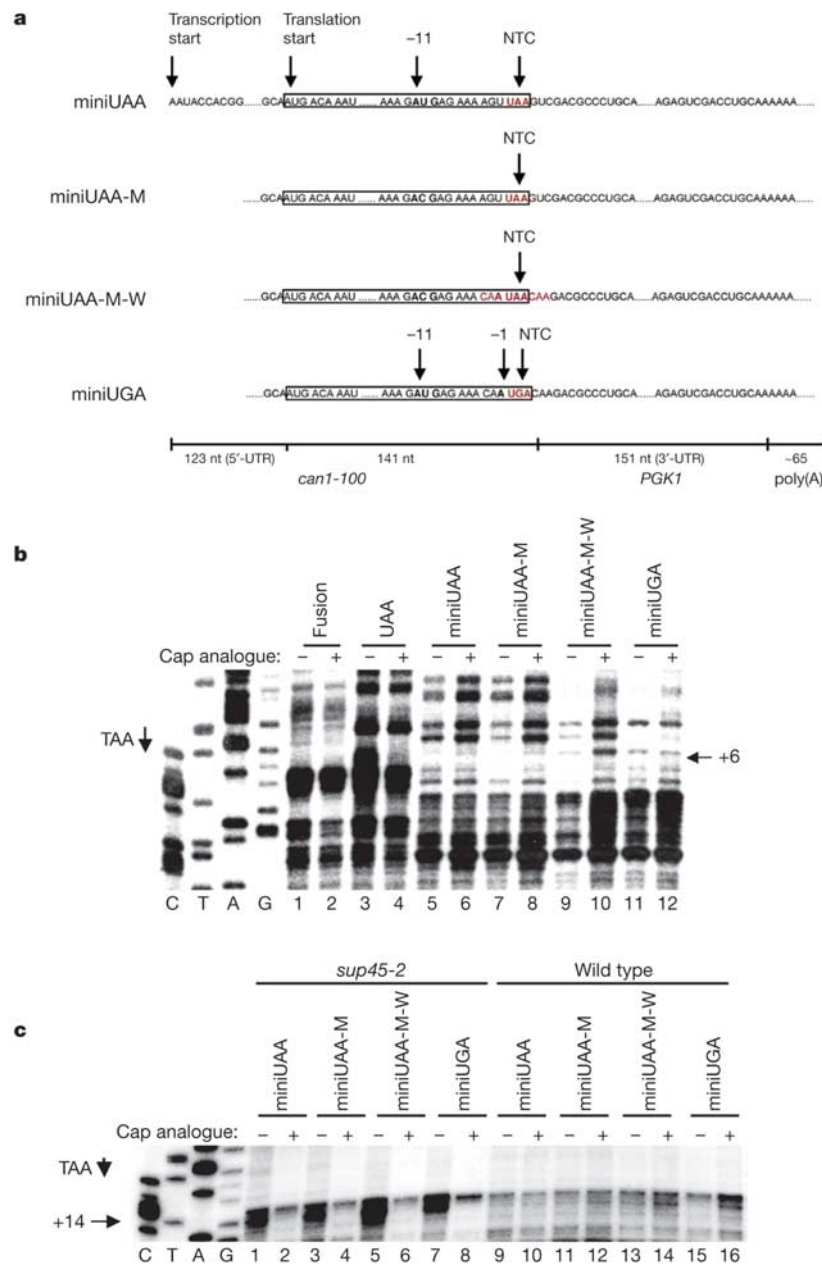


Figure 3 Aberrant toeprint signals are eliminated when premature termination codons (PTCs) are flanked by a normal 3'-UTR. **a**, Pertinent sequences of the 'mini' RNAs; NTC denotes a PTC that has been converted to a normal termination codon by virtue of coding region deletions. Termination codons, and flanking sequences that have been mutated, are indicated in red. **b**, The +6 toeprints are not detected during the translation of 'mini'

RNAs in wild-type extracts in the presence of cycloheximide. The UAA and Fusion RNAs, used as controls, were toeprinted by using oligoprimers no. 3029. **c**, Translation of 'mini' RNAs in the absence of cycloheximide, in *sup45-2* extracts but not in wild-type extracts yields +14 toeprints. The left portions of **b** and **c** show dideoxynucleotide sequencing reactions for the miniUAA template.

the presence of MS2–Pab1p or MS2–Sxl, this method selectively precipitates *PGK1* nonsense mRNAs containing the MS2 binding site (Fig. 4a, lanes 4 and 7). Immunoprecipitation of these mRNPs from extracts containing MS2–Pab1p led to the recovery of MS2–Pab1p and Sup35p (Fig. 4b, left panel), but did not lead to the recovery of the eIF4G or eIF4E initiation factors (data not shown). We attribute co-immunoprecipitation of Sup35p to a specific interaction with MS2–Pab1p because the reciprocal co-immunoprecipitation was also productive (Fig. 4b, right panel) and because no comparable recovery of Sup35p was obtained from immunoprecipitates of MS2–Pab1p lacking the Sup35p-interaction domain³⁰ (Supplementary Fig. 4e) or from immunoprecipitates of *PGK1* nonsense-containing mRNPs stabilized by the deletion of *UPF1* in MS2–Sxl-expressing cells (Supplementary Fig. 4f). Recovery of Sup35p, a termination factor known to interact with Pab1p and

thought to have a function in the stability of conventional mRNAs^{22,23}, indicated a role for this protein in the tethered-Pab1p-mediated reversal of NMD. This seems to be so, because tethered Sup35p also stabilized *PGK1* nonsense transcripts, although to a smaller extent than tethered Pab1p (Fig. 4c). Similar experiments in which Sup45p was the tethered component failed to stabilize the same mRNA significantly, indicating that regulatory aspects of termination had a key function in antagonizing NMD and that Pab1p's role in this process most probably reflected its function as a scaffold for post-transcriptional regulators²¹.

Our results indicate that not all termination events are equivalent and that, at least in yeast, NMD is triggered by a ribosome's failure to terminate adjacent to a properly configured 3'-UTR. As suggested by the *faux* UTR model^{5,8}, proper termination of translation and normal rates of mRNA decay are likely to require interactions

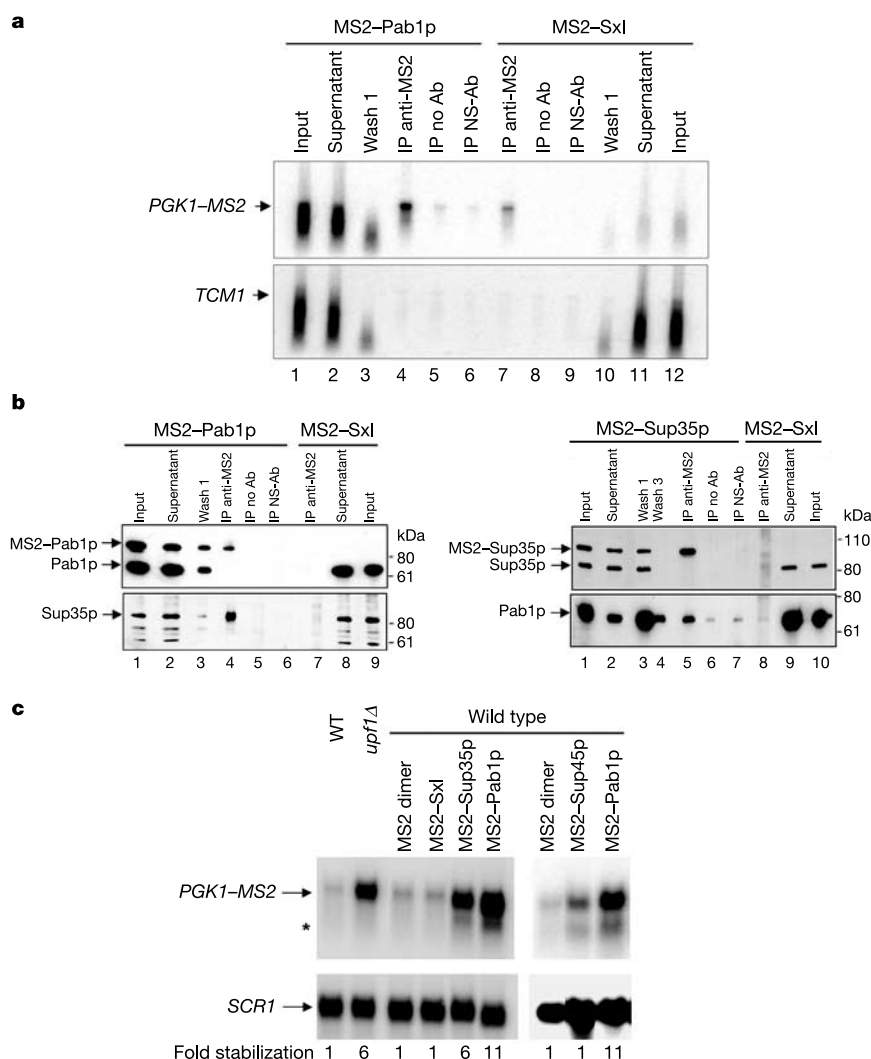


Figure 4 Stabilization of nonsense-containing mRNAs by tethered Pab1p. **a**, *PGK1–MS2* mRNA is selectively recovered in anti-MS2 immunoprecipitates. RNA extracted from immunoprecipitation (IP) reactions with anti-MS2 coat protein antibodies (anti-MS2), no antibodies (no Ab), or non-specific antibodies (NS–Ab) was analysed for the presence of *PGK1–MS2* mRNA or control *TCM1* mRNA. *PGK1–MS2* mRNA is not stabilized in cells expressing MS2–Sxl fusion protein (lanes 7–12), thus rendering its signal weaker than in MS2–Pab1p-expressing cells (lanes 1–6). The loading for the input, the unbound fraction (supernatant) and the wash represents half of the loading of the different IP fractions. **b**, Sup35p co-immunoprecipitates with tethered MS2–Pab1p and vice versa. Left: western blot analysis of IP reactions described in **a** with antibodies that detect Pab1p (upper panel)

or Sup35p (lower panel). Right: lysates from cells expressing *PGK1–MS2* and MS2–Sup35p were immunoprecipitated with anti-MS2 antibodies and analysed for the presence of Sup35p (upper panel) or Pab1p (lower panel). The positions of Pab1p, MS2–Pab1p, Sup35p, MS2–Sup35p and marker proteins are indicated. **c**, mRNA containing a premature termination codon is stabilized by tethered Sup35p but not by tethered Sup45p. RNA extracted from wild-type or *upf1Δ* cells expressing *PGK1–MS2* mRNA was analysed by northern blotting. Cells used for the wild-type (WT) and *upf1Δ* samples contained no fusion protein, whereas the WT cells of the remaining lanes contained the indicated MS2 fusion proteins. The fold stabilization, relative to the WT sample, was determined by phosphorimaging.

between a terminating ribosome and a specific RNP structure or set of factors (including Pab1p and Sup35p) localized 3' to the stop codon. Failure to terminate in this manner seems to slow ribosome release and facilitate reinitiation. The latter event must only be a symptom of aberrant termination, not a cause of NMD, because elimination of reinitiation site AUG codons does not alter the decay rate of nonsense-containing mRNAs (data not shown). However, the failure to release a terminating ribosome effectively might be a stimulus for NMD, possibly promoted by the ribosome's failure to recycle *in cis*²³ and/or by the recruitment of Upf1p to the aberrant termination event or to a bound negative regulatory factor. At a minimum, the absence of aberrant termination toeprints in extracts prepared from *upf1Δ* cells links aberrant termination to NMD and indicates that inactivation of this factor might preclude or perturb events that regulate ribosome stability and movement at a premature stop codon. Finally, we note that it is unlikely that the mechanism of NMD is fundamentally different in yeast and metazoans. If aberrant termination is also the trigger for NMD in higher organisms, then exon junction complexes (EJCs) deposited during splicing of mammalian pre-mRNAs might not reflect a specialized mRNA decay apparatus poised for activation in an early round of translation^{6,7,24–26}. Rather, EJC factors might function like the proteins associated with a normal 3'-UTR in yeast and render an mRNA competent for proper translation when present in an appropriate context. □

Methods

Synthetic mRNAs

Synthetic, capped poly(A)-containing RNA was synthesized *in vitro* from chimaeric genes cloned in a pSP65A vector that included about 65 dT residues for transcription of a poly(A) tail (Promega). The construct used for the synthesis of UAA RNA contained a 5' fragment of the *Saccharomyces cerevisiae can1-100* allele upstream of the firefly *LUC* ORF. The *can1-100* DNA fragment used for this construct contained *EcoRI* and *Sall* sites, 114 base pairs 5' and 147 base pairs 3' of the initiator ATG codon, respectively. This DNA fragment was generated in a one-step polymerase chain reaction (PCR) amplification, using oligonucleotides no. 13 (5'-CCGGAATTCGGATCCAGTTTAAATCTGTCG-3') and no. 12 (5'-CGTCGACTTAACCTTTCTCATCTTTC-3') and cloned into the *EcoRI* and *Sall* restriction sites of the pSP65A vector. The remainder of the construct was derived by the insertion of a *Sall LUC/PGK1-3'* fragment into the *Sall* restriction site of the same plasmid, generating pSP65A-*can1-100/LUC*. All other constructs used here were derived from pSP65A-*can1-100/LUC* by site-directed mutagenesis (Stratagene). The mini-DNA constructs were generated by deletion of the *LUC* ORF from a *PstI* restriction site (created by site-directed mutagenesis) 10–15 nt downstream of the premature stop codon and a *PstI* site immediately 3' to the normal stop codon. All plasmid constructions were confirmed by DNA sequencing. Capped and polyadenylated RNAs were synthesized with the SP6 mMessage mMachine kit (Ambion), in accordance with the manufacturer's protocol, from *HindIII*-linearized plasmids. RNA yields were quantified by spectrophotometry and their integrities were assessed by agarose-gel electrophoresis.

Extracts and translation reactions

S. cerevisiae strains MBS (*MATa ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100* [rho +] L-o, M-o)²⁷, NA101 (a *upf1::HIS3* derivative of MBS constructed for this study) and MT552/8a (*MATa sup45-2 ade2-1 ura3-1*)¹⁵ were used to make extracts for translation *in vitro* as described¹⁰. Translation assays were as described¹⁰, except that each assay used 120 µg of extract, and the treatment with micrococcal nuclease was performed at 25 °C for 5 min. Unless indicated otherwise, translation reactions were incubated for 4 min with 120 ng of each RNA substrate, and terminated by incubation for 3 min with cycloheximide (0.6 mg ml⁻¹). Samples from the translation reactions were used for reverse transcription reactions with the oligoprimers no. 3029, complementary to the *LUC* sequence 5'-CGCCCTGGTTCCTGGAACAATTGC-3'. In the cycloheximide chase experiment (Fig. 1e), the drug was added after 4 min of translation and the reactions were incubated for 0–5 min. Translation reactions programmed with 'mini' RNAs were incubated with 12 ng of the miniUAA, miniUAA-M, miniUAA-M-W or miniUGA RNAs, and toeprinting analyses of aliquots from each reaction used oligoprimers no. 55 complementary to the *PGK1* 3'-UTR sequence 5'-CCGAAAAGAAATAATTGAATTG-3'. Both primers 3029 and 55 hybridize to their respective RNAs at comparable positions. Translation reactions programmed with *Inf* constructs (Supplementary Fig. 2) were incubated for 45 min with 120 ng of RNA, and luciferase activity was assayed as described previously²⁸. All data reported are representative of at least three independent experiments.

Analysis of mRNAs with MS2 binding sites

PCR-amplified 188-nt MS2 coat protein-binding sites²⁹ were inserted 73 nt 3' of premature termination codons in the *CAN1* UAA and UGA genes (see above) and 37 (*PGK1-MS2*) or 164 (*PGK1-3'-MS2*) nt 3' of the UAG codon of mini-*PGK1* (ref. 19).

These constructs were cloned in pYX142 (*CAN1*) or pRS316 (*PGK1*) and introduced into cells harbouring YCplac111-*MS2-PAB1* or pA-MP2 (ref. 29), YCplac111-*MS2-dimer* or pD-*MS2* (ref. 29), pMS2-Sxl (ref. 29), YCplac111-*MS2-SUP35*, YCplac111-*MS2-SUP45* (with *MS2* fusions under the control of the *TPH1* promoter), or no plasmid. RNA was isolated and analysed by northern blotting¹⁹. Hybridization probes were prepared by random priming of double-stranded DNA. *PGK1-MS2* and *PGK1-3'-MS2* mRNAs were detected with a 0.5-kilobase *BglII* fragment of the *MS2*-binding domain; detection of the *CYH2*, *TCM1* and *SCR1* RNAs used gene-specific PCR fragments; and detection of *CAN1-MS2-LUC* mRNAs used a fragment of the *LUC* ORF.

Immunoprecipitation (IP)

Yeast cells, grown at 30 °C in minimal medium to an absorbance $A_{600\text{ nm}}$ of 1, were harvested, then washed and resuspended in IP buffer (10 mM Tris-HCl pH 7.5, 30 mM MgCl₂, 100 mM NaCl, 2 mM phenylmethylsulphonyl fluoride, 2 × protease inhibitor cocktail (Sigma), 0.01% Triton X-100, 50 µg ml⁻¹ cycloheximide, 0.1 mM dithiothreitol). Glass beads (diameter 425–600 µm) were added and cells were broken by vortex-mixing eight times for 15 s, with intermittent cooling on ice. Lysates were clarified twice by centrifugation (7,700 g at 4 °C). Protein A-Sepharose beads (Pierce) were preincubated for 3 h at 4 °C in IP buffer with rabbit polyclonal anti-*MS2* coat protein antibodies or normal rabbit serum (Pierce), precoated overnight at 4 °C with 1 mg of MBS cell extract in 400 µl of IP buffer, washed three times with 1 ml of IP buffer, mixed with 400-µg aliquots of clarified cell extract, and incubated at 4 °C for 3 h, under agitation. Beads were separated from unbound proteins (supernatant) by centrifugation for 10 s and washed four times with IP buffer at 4 °C. The bound fraction was eluted by being boiled for 5 min in SDS gel sample buffer. Wash samples were precipitated with trichloroacetic acid and resuspended in sample buffer. Samples were fractionated by SDS-polyacrylamide-gel electrophoresis and characterized by western blotting. Input, supernatant and wash represented 1/20, 1/10 and 1/3 of the loading of the IP samples, respectively. RNA recovery from IPs was performed as above, with two changes: RNasin (80 U/ml) was added to the IP buffer and the beads were washed with 250-µl aliquots of this solution. RNA was recovered from the beads and IP fractions by extraction with phenol/chloroform¹⁹.

Received 17 February; accepted 20 September 2004; doi:10.1038/nature03060.

- Jacobson, A. & Peltz, S. W. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem.* **65**, 693–739 (1996).
- Maquat, L. E. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. *RNA* **1**, 453–465 (1995).
- Pulak, R. & Anderson, P. mRNA surveillance by the *Caenorhabditis elegans smg* genes. *Genes Dev.* **7**, 1885–1897 (1993).
- Gonzalez, C. I., Bhattacharya, A., Wang, W. & Peltz, S. W. Nonsense-mediated mRNA decay in *Saccharomyces cerevisiae*. *Gene* **274**, 15–25 (2001).
- Maderazo, A. B., Belk, J. P., He, F. & Jacobson, A. Nonsense-containing mRNAs that accumulate in the absence of a functional nonsense-mediated mRNA decay pathway are destabilized rapidly upon its restitution. *Mol. Cell. Biol.* **23**, 842–851 (2003).
- Gatfield, D., Unterholzner, L., Ciccarelli, F. D., Bork, P. & Izaurralde, E. Nonsense-mediated mRNA decay in *Drosophila*: at the intersection of the yeast and mammalian pathways. *EMBO J.* **22**, 3960–3970 (2003).
- LeBlanc, J. J. & Beemon, K. L. Unspliced Rous sarcoma virus genomic RNAs are translated and subjected to nonsense-mediated mRNA decay before packaging. *J. Virol.* **78**, 5139–5146 (2004).
- Jacobson, A. & Peltz, S. W. in *Translational Control* (eds Sonenberg, N., Hershey, J. W. B. & Mathews, M. B.) 827–847 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2000).
- Hillgren, P. & Parker, R. mRNA surveillance in eukaryotes: kinetic proofreading of proper translation termination as assessed by mRNP domain organization? *RNA* **5**, 711–719 (1999).
- Sachs, M. S. *et al.* Toeprint analysis of the positioning of translation apparatus components at initiation and termination codons of fungal mRNAs. *Methods* **26**, 105–114 (2002).
- Maderazo, A. B., He, F., Mangus, D. A. & Jacobson, A. Upf1p control of nonsense mRNA translation is regulated by Nmd2p and Upf3p. *Mol. Cell. Biol.* **20**, 4591–4603 (2000).
- Bonetti, B., Fu, L., Moon, J. & Bedwell, D. M. The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **251**, 334–345 (1995).
- Dmitriev, S. E., Pisarev, A. V., Rubtsova, M. P., Dunaevsky, Y. E. & Shatsky, I. N. Conversion of 48S translation preinitiation complexes into 80S initiation complexes as revealed by toeprinting. *FEBS Lett.* **533**, 99–104 (2003).
- Kozak, M. Primer extension analysis of eukaryotic ribosome-mRNA complexes. *Nucleic Acids Res.* **26**, 4853–4859 (1998).
- Stansfield, I., Kushnirov, V. V., Jones, K. M. & Tuite, M. F. A conditional-lethal translation termination defect in a *sup45* mutant of the yeast *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **245**, 557–563 (1997).
- Thomas, K. R. & Capecchi, M. R. Introduction of homologous DNA sequences into mammalian cells induces mutations in the cognate gene. *Nature* **324**, 34–38 (1986).
- Peabody, D. S. & Berg, P. Termination-reinitiation occurs in the translation of mammalian cell mRNAs. *Mol. Cell. Biol.* **6**, 2695–2703 (1986).
- Song, H. *et al.* The crystal structure of human eukaryotic release factor eRF1—mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. *Cell* **100**, 311–321 (2000).
- Peltz, S. W., Brown, A. H. & Jacobson, A. mRNA destabilization triggered by premature translational termination depends on at least three *cis*-acting sequence elements and one *trans*-acting factor. *Genes Dev.* **7**, 1737–1754 (1993).
- Muhlrad, D. & Parker, R. Recognition of yeast mRNAs as 'nonsense containing' leads to both inhibition of mRNA translation and mRNA degradation: implications for the control of mRNA decapping. *Mol. Biol. Cell* **10**, 3971–3978 (1999).
- Mangus, D. A., Evans, M. C. & Jacobson, A. Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol.* **4**, 223.1–223.14 (2003).
- Hosoda, N. *et al.* Translation termination factor eRF3 mediates mRNA decay through the regulation of deadenylation. *J. Biol. Chem.* **278**, 38287–38291 (2003).
- Uchida, N., Hoshino, S., Imataka, H., Sonenberg, N. & Katada, T. A novel role of the mammalian

- GSPT/eRF3 associating with poly(A)-binding protein in Cap/Poly(A)-dependent translation. *J. Biol. Chem.* **277**, 50286–50292 (2002).
24. Kim, V. N., Kataoka, N. & Dreyfuss, G. Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon–exon junction complex. *Science* **293**, 1832–1836 (2001).
 25. Gehring, N. H., Neu-Yilik, G., Schell, T., Hentze, M. W. & Kulozik, A. E. Y14 and hUpf3 form an NMD-activating complex. *Mol. Cell* **11**, 939–949 (2003).
 26. Le Hir, H., Izaurralde, E., Maquat, L. E. & Moore, M. J. The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon–exon junctions. *EMBO J.* **19**, 6860–6869 (2000).
 27. Iizuka, N. & Sarnow, P. Translation-competent extracts from *Saccharomyces cerevisiae*: effects of L-A RNA, 5' cap, and 3' poly(A) tail on translational efficiency of mRNAs. *Methods* **11**, 353–360 (1997).
 28. Tarun, S. Z. Jr & Sachs, A. B. A common function for mRNA 5' and 3' ends in translation initiation in yeast. *Genes Dev.* **9**, 2997–3007 (1995).
 29. Collier, J. M., Gray, N. K. & Wickens, M. P. mRNA stabilization by poly(A) binding protein is independent of poly(A) and requires translation. *Genes Dev.* **12**, 3226–3235 (1998).
 30. Cosson, B. *et al.* Poly(A)-binding protein acts in translation termination via eukaryotic release factor 3 interaction and does not influence [PSI(+)] propagation. *Mol. Cell. Biol.* **22**, 3301–3315 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Stockley, University of Leeds, for MS2 coat protein antibodies; T. Serio (Brown University) and D. C. Masison (NIH) for Sup35p antibodies; J. McCarthy (UMIST) for eIF4G and eIF4E antibodies; and M. P. Wickens (University of Wisconsin–Madison) for plasmids carrying MS2 constructs. This work was supported by grants to A.J. from the National Institutes of Health and a postdoctoral fellowship to N.A. from the Human Frontier Science Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.J. (allan.jacobson@umassmed.edu).

Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase

Ralf Thoma, Tanja Schulz-Gasch, Brigitte D'Arcy, Jörg Benz, Johannes Aebi, Henrietta Dehmlow, Michael Hennig, Martine Stihle & Armin Ruf

F. Hoffmann-La Roche AG, Pharma Research Discovery Chemistry, 4070 Basel, Switzerland

In higher organisms the formation of the steroid scaffold is catalysed exclusively by the membrane-bound oxidosqualene cyclase (OSC; lanosterol synthase). In a highly selective cyclization reaction OSC forms lanosterol with seven chiral centres starting from the linear substrate 2,3-oxidosqualene. Valuable data on the mechanism of the complex cyclization cascade have been collected during the past 50 years using suicide inhibitors, mutagenesis studies and homology modelling. Nevertheless it is still not fully understood how the enzyme catalyses the reaction^{1,2}. Because of the decisive role of OSC in cholesterol biosynthesis it represents a target for the discovery of novel anticholesteremic drugs that could complement the widely used statins³. Here we present two crystal structures of the human membrane protein OSC: the target protein with an inhibitor that showed cholesterol lowering *in vivo* opens the way for the structure-based design of new OSC inhibitors. The complex with the reaction product lanosterol gives a clear picture of the way in which the enzyme achieves product specificity in this highly exothermic cyclization reaction.

Human OSC consists of two (α/α) barrel domains that are connected by loops and three smaller β -structures. The large active-site cavity is located in the centre of the molecule between domains 1 and 2 (Fig. 1). The amino-terminal OSC region, which is

absent from other cyclase sequences, fills the space between the two domains and could function in stabilizing their relative orientations (Fig. 1). Five cyclase fingerprint QW-sequence motifs are located towards the carboxy terminus from outer barrel helices. They assume the same conformation with stacked Gln and Trp side chains as that observed in the bacterial squalene hopene cyclase (SHC); the conformation is thought to contribute to cyclase fold stability during the highly exergonic cyclization reaction^{4,5}. Human OSC is monomeric in the crystal, which is consistent with analytical ultracentrifuge data⁶.

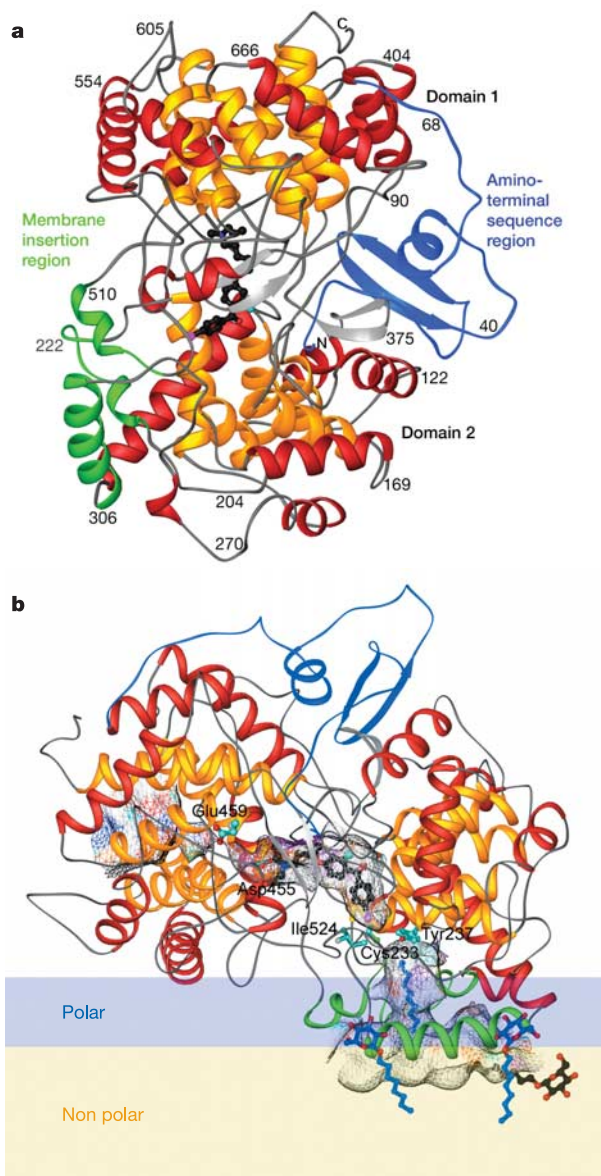


Figure 1 Ribbon diagram of human OSC. **a**, The C and N termini and several sequence positions are labelled. The inner barrel helices are coloured yellow. The bound inhibitor (black) indicates the location of the active site. **b**, The orientation of OSC relative to one leaflet of the membrane, whose polar and nonpolar parts are depicted in light blue and light yellow respectively. Internal surfaces and channels of OSC are shown with the following colour code: blue, positive; red, negative; cyan, hydrogen-bond donor; magenta, other polar. Ro 48-8071 binds in the central active-site cavity. Two channels lead to the enzyme surface: one is hydrophobic to the membrane insertion site and one is polar. The fragment of lipid (blue) binds to the hydrophobic substrate entrance channel. A β -OG molecule belonging to a crystal neighbour (black) interacts with the membrane-inserting hydrophobic surface.

OSC is a monotopic membrane protein; that is, it inserts into the membrane but does not span the bilayer⁷. The OSC membrane-binding region in domain 2 can be detected by bound detergent molecules and by its hydrophobic surface ($1.700 \text{ \AA}^2$), which makes up 6% of the total enzyme surface (Fig. 1). The membrane-inserted surface consists of a plateau $25 \text{ \AA}$ in diameter and a channel that leads to the active-site cavity. This channel is supposed to allow the substrate oxidosqualene to enter the hydrophobic active site but a constriction site separates it from the active-site cavity (Fig. 1b)^{4,8}. Passage of the substrate could be achieved either by a change in the sidechain conformations of the residues Tyr 237, Cys 233 and Ile 524 or by rearrangement of the strained loops 516–524 and 697–699 that contain the only two residues in the disallowed regions of the

Ramachandran plot (Glu 519 and Lys 698). The electron density observed in the substrate entrance channel can be fitted with a tetradecyl carbon chain but not with the octyl chain of the detergent octyl- β -D-glucopyranoside (β -OG) and might belong to a partly ordered lipid (Fig. 1b). In the crystals of other monotopic membrane enzymes, detergents have been observed in the hydrophobic substrate entrance channels^{9,10}. The approximate orientation of OSC in the membrane was inferred by aligning the glucose headgroups of two bound β -OG detergent molecules in the polar membrane layer and the octyl groups parallel to the fatty-acid chains in the bilayer (Fig. 1b). The two polar headgroups are bound along one of the three helices that constitute the membrane-binding region (Fig. 1b). The nonpolar octyl tail of one β -OG molecule interacts with the hydrophobic region of a neighbouring OSC molecule in a crystal lattice contact. The defined binding and involvement in a crystal lattice contact might explain why β -OG was the only detergent that yielded OSC crystals⁶.

The catalytic mechanism for the polycyclization reaction involves several reaction steps (Fig. 2)². First, (3S)-2,3-oxidosqualene adopts a pre-organized chair–boat–chair conformation. Protonation of the epoxide ring then triggers a cascade of ring-forming reactions to the protosterol cation. Skeletal rearrangement of this intermediate through a series of 1,2-hydride and 1,2-methyl group shifts and a final deprotonation step leads to the product lanosterol². Product specificity and high stereoselectivity are believed to be achieved by several factors: first, forcing the substrate to occupy a prefolded conformation, second progression of the reaction through rigidly held, partly cyclized carbocationic intermediates, and last, stabilization of the intermediate carbocations by cation– π interactions^{11–14}, thus preventing early truncation of the cyclization cascade by deprotonation or by the nucleophilic addition of solvent molecules.

The homologue structure of SHC is only a limited model for understanding the OSC reaction, because SHC differs in the substrate squalene, in the activation of the catalytic acid by a histidine residue, in the B-ring chair conformation, the expanded D-ring, and the additional E-ring of its product hopene, and in the missing final rearrangement step^{1,15,16}. To gain more insight into the structural basis for this highly stereoselective cyclization, we crystallized OSC together with its product lanosterol. The ligand was well defined in the difference map at $2.1 \text{ \AA}$ resolution, with clear electron density for the sterol scaffold and all lanosterol methyl substituents (Fig. 3a). Lanosterol fits closely to the shape and physicochemical properties of the OSC active-site cavity. One hydrogen bond is formed between the lanosterol 3-hydroxy group and the catalytic Asp 455 that constitutes the polar cap of the mainly hydrophobic cavity. On the basis of this crystal structure the 2,3-oxidosqualene cyclization intermediates were modelled into the active site. This confirmed the hypothesis that initiation of the cyclization reaction is accomplished by Asp 455 protonating the epoxide group of prefolded 2,3-oxidosqualene (Fig. 2). 2,3-Oxidosqualene is completely stable in glacial acetic acid for about 1 day, so Asp 455 is expected to be further activated¹⁷. The X-ray structure assigns Cys 456 and Cys 533 to act as hydrogen-bonding partners with Asp 455 (Fig. 3) and thus to contribute to the required increased acidity of Asp 455 in OSC¹⁸. Reprotonation of Asp 455 after the completion of one catalytic cycle can be achieved either from the bulk solvent through a chain of water molecules and the carboxylate group of Glu 459 (Fig. 1) or by shifting the proton from the final deprotonation step back to Asp 455.

The conserved side-chains of Phe 444, Tyr 503 and Trp 581 are in an appropriate position and orientation to stabilize the intermediate tertiary cations at C-6 and C-10, after A-ring and B-ring formation, through cation– π interactions (Fig. 3a)¹. Tyr 98 is well positioned to enforce the energetically unfavourable boat conformation of 2,3-oxidosqualene required for lanosterol B-ring

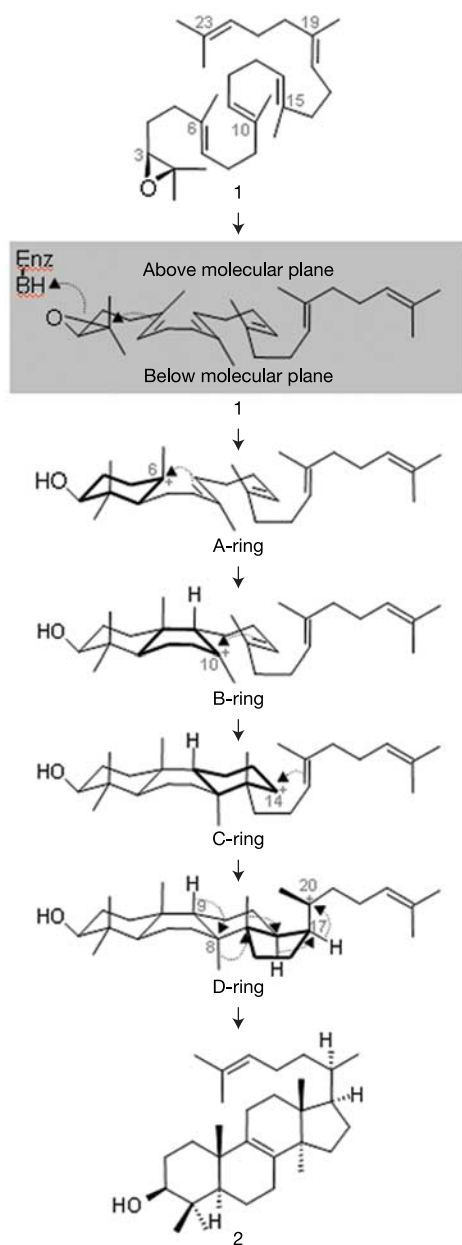


Figure 2 OSC catalyses the conversion of 2,3-oxidosqualene **1** to lanosterol **2**. The initial substrate chair–boat–chair conformation of 2,3-oxidosqualene **1**, putative conformations after A-ring, B-ring and C-ring closure (2,3-oxidosqualene cation numbering), and skeletal rearrangement of protosterol cation (lanosterol cation numbering) through 1,2-shifts of hydride and methyl groups are shown.

formation by pushing the methyl group at C-10 below the molecular plane (Figs 2 and 3b). In the OSC sequence this is reflected by a one-residue insertion above and a one-residue deletion below the molecular plane as opposed to SHC, which creates the B-ring in the chair conformation¹⁵. His 232 and Phe 696 side chains are positioned to stabilize the anti-Markovnikov secondary cation created at C-14 during C-ring formation (Fig. 3), through π interactions. The OSC cyclization cascade stops with the tertiary

protosterol cation at C-20 after formation of the five-membered D-ring (Fig. 2), because OSC lacks an aromatic residue like Trp 169 in SHC, which stabilizes the long-lived secondary cation at C-17 (protosterol cation numbering) required for hopene E-ring cyclization^{15,16}.

The protosterol C-20 cation is converted to the lanosterol C-8/C-9 cation by skeletal rearrangement. The equilibrium between these cations is shifted, because the active site has a higher

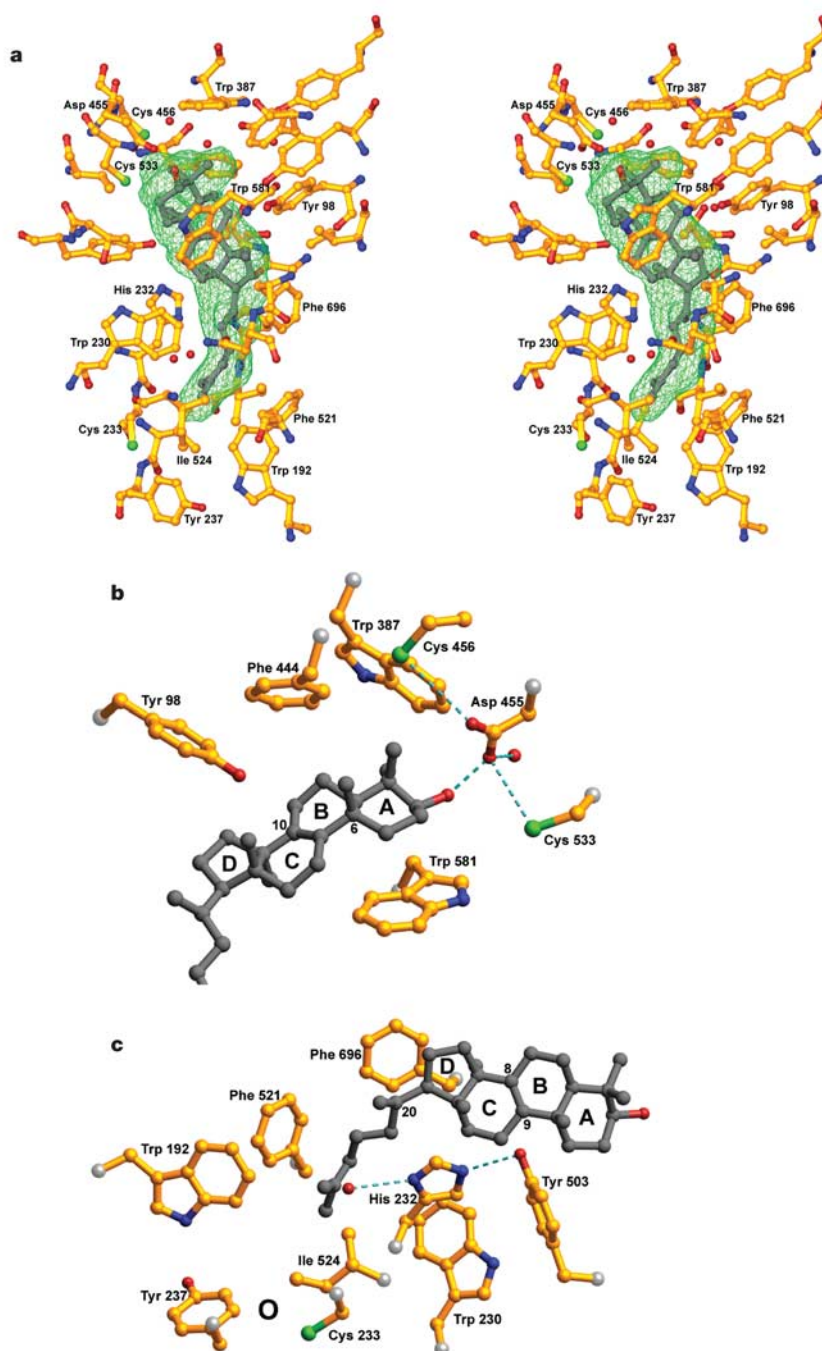


Figure 3 Cyclization mechanism in the light of the OSC-lanosterol complex structure. **a**, Stereo view of the autoBUSTER difference electron density contoured at 3σ. Residues at a distance less than 5 Å are shown. Water molecules are observed only near Asp 455 and His 232. **b**, A- and B-rings: cation- π interactions of Trp 387, Phe 444 and Trp 581 can stabilize the cyclization intermediates positively charged at oxidosqualene positions 6

and 10. The catalytic acid Asp 455 is activated by Cys 456 and Cys 533. The Tyr 98 side chain sterically hinders the B-ring from assuming the favourable chair conformation. **c**, C- and D-rings: Phe 696 and the His 232 can stabilize the positive charge at the C-20 protosterol cation by cation- π interactions. His 232 is the nearest basic residue that could deprotonate the C-8/C-9 lanosterol cation.

π -electron density near C-8/C-9 with seven aromatic residues, in comparison with three residues in a distance of 6 Å to C-20. Truncation of the reaction by the nucleophilic addition of water molecules is avoided by the predominantly hydrophobic nature of the active-site cavity (Fig. 3). In addition, the absence of basic residues in the proximity of the charge of the protosterol cation avoids premature truncation by deprotonation. As predicted in previous studies^{15,19}, the highly conserved His 232 is the only basic residue close enough to accept the proton in the specific deprotonation of the C-8/C-9 lanosterol cation that terminates catalysis (Fig. 3c). However, the hydroxy group of Tyr 503 that is hydrogen-bonded to His 232 would be in a proper position for the final deprotonation step.

Inhibitors of OSC as anticholesteremic drugs act downstream of farnesyl-pyrophosphate in the cholesterol pathway and do not interfere with the synthesis of isoprenoids and coenzyme Q. In hamsters, pharmacologically active doses of an OSC inhibitor showed less adverse effects than a statin³. The structure of the benchmark inhibitor Ro 48-8071 (Fig. 4) in complex with human OSC provides a structural base for the design of improved OSC inhibitors. A single molecule of the co-crystallized inhibitor

Ro 48-8071 was well defined in the initial 2.2 Å ($F_o - F_c$) electron density map. This is in agreement with competitive inhibitor binding and the 1:1 binding ratio derived from a fluorescence titration experiment^{6,20}.

The basic nitrogen atom of the inhibitor Ro 48-8071 forms a charged hydrogen bond at a distance of 2.9 Å from the Asp 455 carboxylate that directs both oxygen atoms towards the active-site cavity. The charged amino group is also stabilized through cation- π interactions with several aromatic residues (Fig. 4a). The fluoro-phenyl group of the inhibitor benzophenone moiety is stacked between the side chains of Phe 696 (3.7 Å) and His 232 (3.1 Å). The carbonyl group is hydrogen-bonded to a water molecule, which in turn interacts with the backbone amide nitrogen of Ile 338 (Fig. 4b). This water-mediated interaction explains why the hydrogen-bond donor or acceptor properties of groups substituting for benzophenone have no effect on the concentration giving half-maximal inhibition²¹. The terminal phenyl group interacts with Trp 192 (3.6 Å), Trp 230 (3.7 Å) and Phe 521 (4.6 Å) (Fig. 4b). The π -electron-rich pocket created by these residues is optimal for the stabilization of electron-deficient aromatic systems²². The bromine atom of the inhibitor is interacting with the residues Ile 524, Tyr 237 and Cys 233 close to the channel constriction site. The length of the linker between the amine nitrogen and the electron-deficient aromatic group has a pronounced effect on inhibitor binding affinity²³. The structure pinpoints the fact that linkers that are too long lead to steric clashes or constrained inhibitor conformations, and short inhibitors do not fill the binding site completely and lose binding affinity. The direct hydrogen bond between the catalytic Asp 455 and the amino group of Ro 48-8071 is not observed in SHC²⁴. In addition, the OSC active-site cavity is smaller than that of SHC, which has to harbour the additional hopene E-ring. This leads to different inhibitor conformations in SHC and OSC and explains why inhibitor design based on SHC resulted in inaccurate predictions for OSC. Our co-crystal structure prompted the design of novel structural classes of OSC inhibitors, such as amides that lack the charged amino group and therefore possess improved physico-chemical properties.

The high-resolution crystal structures of human OSC provide for the first time a clear picture of the molecular mechanism of substrate entry and stereoselective cyclization reaction of sterol formation. The structure confirms the assignment of Asp 455 as the catalytic acid, its further increase in acidity by the two cysteines Cys 456 and Cys 533, and His 232 as the basic residue involved in termination of the cyclization reaction. Further biochemical investigation should be directed to confirming the proposed function of the hydrophobic substrate channel and aromatic residues that are crucial for the π interaction with the reaction intermediates. Drug design will be facilitated by the elucidation of key interaction that can be used for the improvement of the inhibitory and physico-chemical properties of drug molecules for decreasing cholesterol concentration in humans. □

Methods

Recombinant human OSC in a buffer containing 0.8% (w/v) β -OG was used for crystallization as described⁶. For inhibitor co-crystals the protein was incubated overnight with 5 mM Ro 48-8071. Crystals were grown at 298 K in hanging drops by mixing equal volumes of protein solution with the reservoir solution consisting of 25% (w/v) PEG 3350, 0.4 M ammonium acetate, 0.1 M Tris-HCl pH 8.5, 10% (v/v) ethylene glycol and the addition of seed crystals. Untwinned crystals were selected for data collection on the beamline X06SA (Swiss Light Source). Data were processed with Denzo²⁵. The structure was solved by molecular replacement with SHC¹⁰ (Protein Data Bank entry 2sqc) as search model with the use of the CCP4 (ref. 26) program suite and was rebuilt by ARP/wARP²⁷. After addition of the inhibitor and detergents with MOLOC²⁸, the model was refined with REFMAC²⁹. The lanosterol structure was refined with BUSTER³⁰. Difference electron-density maps calculated with autoBUSTER³⁰ from a model refined without a ligand clearly showed at the 3 σ level the sterol scaffold featuring double substituents at C-4 (Fig. 3a). After refinement, the electron density for the B- and C-rings of lanosterol remains weaker than for the rest of the molecule. Data set and refinement statistics are given in Supplementary Table 1. Modelling of substrates and intermediates to the human OSC structure was done with MOE 2003.02 and MOLOC²⁸.

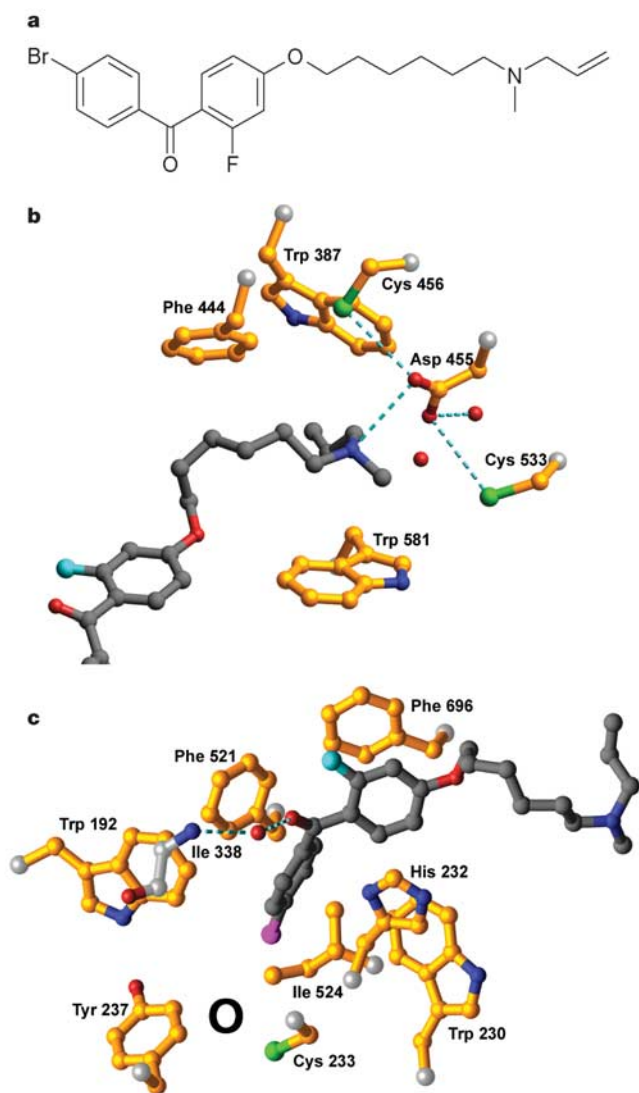


Figure 4 OSC residues interacting with the inhibitor Ro 48-8071 in the crystal structure. **a**, Ro 48-8071. **b**, Interactions with the amino group. **c**, Interactions with the benzophenone group. The constriction site is marked by O.

Received 19 June; accepted 2 September 2004; doi:10.1038/nature02993.

- Wendt, K. U., Schulz, G. E., Corey, E. J. & Liu, D. R. Enzyme mechanisms for polycyclic triterpene formation. *Angew. Chem. Int. Edn Engl.* **39**, 2812–2833 (2000).
- Hoshino, T. & Sato, T. Squalene-hopene cyclase: catalytic mechanism and substrate recognition. *Chem. Commun. (Camb.)* **4**, 291–301 (2002).
- Morand, O. H. *et al.* Ro48–8071, a new 2,3-oxidosqualene:lanosterol cyclase inhibitor lowering plasma cholesterol in hamsters, squirrel monkeys, and minipigs: comparison to simvastatin. *J. Lipid Res.* **38**, 373–390 (1997).
- Wendt, K. U., Poralla, K. & Schulz, G. E. Structure and function of a squalene cyclase. *Science* **277**, 1811–1815 (1997).
- Poralla, K. *et al.* A specific amino acid repeat in squalene and oxidosqualene cyclases. *Trends Biochem. Sci.* **19**, 157–158 (1994).
- Ruf, A. *et al.* The monotopic membrane protein human oxidosqualene cyclase is active as monomer. *Biochem. Biophys. Res. Commun.* **315**, 247–254 (2004).
- Blobel, G. Intracellular protein topogenesis. *Proc. Natl Acad. Sci. USA* **77**, 1496–1500 (1980).
- Milla, P. *et al.* Thiol-modifying inhibitors for understanding squalene cyclase function. *Eur. J. Biochem.* **269**, 2108–2116 (2002).
- Loll, P. J., Picot, D. & Garavito, R. M. The structural basis of aspirin activity inferred from the crystal structure of inactivated prostaglandin H2 synthase. *Nature Struct. Biol.* **2**, 637–643 (1995).
- Wendt, K. U., Lenhart, A. & Schulz, G. E. The structure of the membrane protein squalene-hopene cyclase at 2.0 Å resolution. *J. Mol. Biol.* **286**, 175–187 (1999).
- Corey, E. J., Matsuda, S. P. T., Baker, C. H., Ting, A. Y. & Cheng, H. Molecular cloning of a *Schizosaccharomyces pombe* cDNA encoding lanosterol synthase and investigation of conserved tryptophan residues. *Biochem. Biophys. Res. Commun.* **219**, 327–331 (1996).
- Zoltewicz, J. A., Maier, N. M. & Fabian, W. M. F. π -cation and π -dipole-stabilizing interactions in a simple model system with cofacial aromatic rings. *J. Org. Chem.* **63**, 4985–4990 (1998).
- Miklis, P. C., Ditchfield, R. & Spencer, T. A. Carbocation- π interaction: computational study of complexation of methyl cation with benzene and comparisons with related systems. *J. Am. Chem. Soc.* **120**, 10482–10489 (1998).
- Gallivan, J. P. & Dougherty, D. A. Cation- π interactions in structural biology. *Proc. Natl Acad. Sci. USA* **96**, 9459–9464 (1999).
- Schulz-Gasch, T. & Stahl, M. Mechanistic insights into oxidosqualene cyclizations through homology modeling. *J. Comput. Chem.* **24**, 741–753 (2003).
- Reinert, D. J., Balliano, G. & Schulz, G. E. Conversion of squalene to the pentacarbocyclic hopene. *Chem. Biol.* **11**, 121–126 (2004).
- Corey, E. J. *et al.* Methodology for the preparation of pure recombinant *S. cerevisiae* lanosterol synthase using a baculovirus expression system. Evidence that oxirane cleavage and A-ring formation are concerted in the biosynthesis of lanosterol from 2,3-oxidosqualene. *J. Am. Chem. Soc.* **119**, 1277–1288 (1997).
- Gandour, R. D. On the importance of orientation in general base catalysis by carboxylate. *Bioorg. Chem.* **10**, 169–176 (1981).
- Corey, E. J. *et al.* Studies on the substrate binding segments and catalytic action of lanosterol synthase. Affinity labeling with carbocations derived from mechanism-based analogs of 2,3-oxidosqualene and site-directed mutagenesis probes. *J. Am. Chem. Soc.* **119**, 1289–1290 (1997).
- Abe, I., Zheng, Y. F. & Prestwich, G. D. Photoaffinity labeling of oxidosqualene cyclase and squalene cyclase by a benzophenone-containing inhibitor. *Biochemistry* **37**, 5779–5784 (1998).
- Brown, G. R. *et al.* Quinuclidine inhibitors of 2,3-oxidosqualene cyclase-lanosterol synthase: optimization from lipid profiles. *J. Med. Chem.* **42**, 1306–1311 (1999).
- Dehmlo, H. *et al.* Synthesis and structure-activity studies of novel orally active non-terpenoid 2,3-oxidosqualene cyclase inhibitors. *J. Med. Chem.* **46**, 3354–3370 (2003).
- Jolidon, S., Polak-Wyss, A., Hartmann, P. G. & Guerry, P. In *Recent Advances in the Chemistry of Anti-infective Agents* (ed. Bentley, P. H.) 223–233 (Royal Soc. Chemistry, Cambridge, 1993).
- Lenhart, A., Weihofen, W. A., Pleschke, A. E. & Schulz, G. E. Crystal structure of a squalene cyclase in complex with the potential anticholesteremic drug Ro48–8071. *Chem. Biol.* **9**, 639–645 (2002).
- Otwinowski, Z. & Minor, W. M. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Vagin, A. & Teplov, A. MOLREP: an automated program for molecular replacement. *J. Appl. Crystallogr.* **30**, 1022–1025 (1997).
- Morris, R. J., Perrakis, A. & Lamzin, V. S. ARP/wARP's model-building algorithms. I. The main chain. *Acta Crystallogr. D* **58**, 968–975 (2002).
- Gerber, P. R. Peptide mechanics: a force field for peptides and proteins working with entire residues as small units. *Biopolymers* **32**, 1003–1017 (1992).
- Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. Efficient anisotropic refinement of macromolecular structures using FFT. *Acta Crystallogr. D* **55**, 247–255 (1999).
- Roversi, P., Blanc, E., Vornrhein, C., Evans, G. & Bricogne, G. Modelling prior distributions of atoms for macromolecular refinement and completion. *Acta Crystallogr. D* **56**, 1316–1323 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the staff at the beamline X06SA at the Swiss Light Source (SLS, Switzerland) for support, and C. Vornrhein (Global Phasing Ltd) for an early version of autoBUSTER. We thank all colleagues at Roche Basel for support, and especially O. Morand for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.R. (armin.ruf@roche.com). The atomic coordinates have been deposited at the Protein Data Bank under the accession codes 1w6j (Ro48–8071) and 1w6k (lanosterol).

Membrane structure and interactions with protein and DNA in bacteriophage PRD1

Joseph J. B. Cockburn^{1,2}, Nicola G. A. Abrescia¹, Jonathan M. Grimes¹, Geoffrey C. Sutton¹, Jonathan M. Diprose¹, James M. Benevides³, George J. Thomas Jr³, Jaana K. H. Bamford⁴, Dennis H. Bamford⁴ & David I. Stuart^{1,2}

¹Division of Structural Biology, The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford OX3 7BN, UK

²Oxford Centre for Molecular Sciences, Central Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QT, UK

³Division of Cell Biology and Biophysics, School of Biological Sciences, University of Missouri – Kansas City, Kansas City, Missouri 64110, USA

⁴Institute of Biotechnology and Faculty of Biosciences, Viikki Biocenter, University of Helsinki, P.O. Box 56, Viikinkaari 5, 00014, Finland

Membranes are essential for selectively controlling the passage of molecules in and out of cells and mediating the response of cells to their environment. Biological membranes and their associated proteins present considerable difficulties for structural analysis. Although enveloped viruses have been imaged at about 9 Å resolution by cryo-electron microscopy and image reconstruction^{1,2}, no detailed crystallographic structure of a membrane system has been described. The structure of the bacteriophage PRD1 particle, determined by X-ray crystallography at about 4 Å resolution, allows the first detailed analysis of a membrane-containing virus³. The architecture of the viral capsid and its implications for virus assembly are presented in the accompanying paper³. Here we show that the electron density also reveals the icosahedral lipid bilayer, beneath the protein capsid, enveloping the viral DNA. The viral membrane contains about 26,000 lipid molecules asymmetrically distributed between the membrane leaflets. The inner leaflet is composed predominantly of zwitterionic phosphatidylethanolamine molecules, facilitating a very close interaction with the viral DNA, which we estimate to be packaged to a pressure of about 45 atm, factors that are likely to be important during membrane-mediated DNA translocation into the host cell. In contrast, the outer leaflet is enriched in phosphatidylglycerol and cardiolipin, which show a marked lateral segregation within the icosahedral asymmetric unit. In addition, the lipid headgroups show a surprising degree of order.

PRD1 (family Tectiviridae) possesses a host-derived membrane vesicle beneath its capsid, encapsulating the 15,000-base-pair double-stranded DNA (dsDNA) genome⁴. Trimers of protein P3 form the flat facets of the icosahedral capsid^{5,6}, which are locked together by a ‘cementing’ protein P30 (ref. 7), and spike complexes of P31, P5 and P2 form the vertices⁸. A unique vertex contains the DNA-packaging machinery (ATPase P9 and protein P6), anchored into the viral membrane through association with integral membrane proteins P20 and P22 (refs 9, 10). The viral membrane contains further integral membrane proteins (P7, P11, P14, P16, P18, P32 and P34) and a lytic enzyme, P15, associated with its periphery^{11–15}. It has been reported that the total mass of the membrane associated proteins is about equal to that attributable to lipid¹⁶, and analysis of the transmembrane helix content of these proteins suggests that protein accounts for about 7% of the total mass within the confines of the bilayer. The membrane has a crucial function in the infection process: extrusion of DNA from the virus is associated with transformation of the membrane into a tubular structure protruding from one vertex^{11,17}, which might penetrate the host cell wall (with protein P7 digesting the cellular peptidoglycan)

to form a channel through which the phage genome enters the host cell¹⁵.

Structure determination³ of the Sus539 mutant¹⁷ PRD1 virion produced electron density maps showing concentric layers of electron density beneath the viral capsid (Fig. 1a), the outermost two layers belonging to the lipid headgroups and those beneath to the genome. The electron density lying between the headgroup regions is lower than that of the bulk solvent, as expected for the hydrophobic lipid acyl chains. The membrane is icosahedral, measuring 535 Å between opposing vertices (Fig. 1b). This requires a remarkable variation in the membrane curvature over the icosahedral asymmetric portion. The greatest curvature is induced at the capsid vertices (radius about 75 Å) by the viral membrane protein P16 (ref. 3). Furthermore, a series of interactions anchor the major capsid protein P3 in the viral membrane. Half of the P3 subunits in the icosahedral asymmetric unit have amino-terminal extensions reaching down to the membrane, immersing the first five, disordered, residues in the outer-leaflet headgroup region. These 12 sites of interaction lie along the edges of each icosahedral facet and may stabilize the membrane curvature along the facet edge by packing extra material into the headgroup region of the outer leaflet (Fig. 1c).

The electron density maps indicated lateral order within the membrane (Fig. 1d), despite confirmation from Raman spectroscopic studies that the lipids in the virus crystals were in the liquid-crystalline phase (see Supplementary Methods and Supplementary Fig. 1), as for the virus in solution¹⁸. Careful analysis of the apparent lateral order (see Supplementary Fig. 2) suggested that some of the details of this structure might be spurious; nevertheless, the clear density variations within the headgroup region are reinforced in the crystallographic cell form with the most complete data set and show some correlation between two independent cell forms. Although we have not attempted to produce a detailed atomic model for the membrane, we have picked electron density peaks to make quantitative observations about the distribution of lipid headgroups (see Supplementary Fig. 3). The autocorrelation function for these experimentally derived peaks suggests an 8.5 Å average separation of the lipid headgroups, closely matching that derived for a model membrane.

The lipid composition of the viral membrane for PRD1 (grown in *Salmonella enterica* DS88)¹⁹ is summarized in Supplementary Fig. 4. The lipid headgroups are predominantly phosphatidylethanolamine (PE) and phosphatidylglycerol (comprising 53% and 43% by number, respectively), with 4% cardiolipin. At full hydration (a

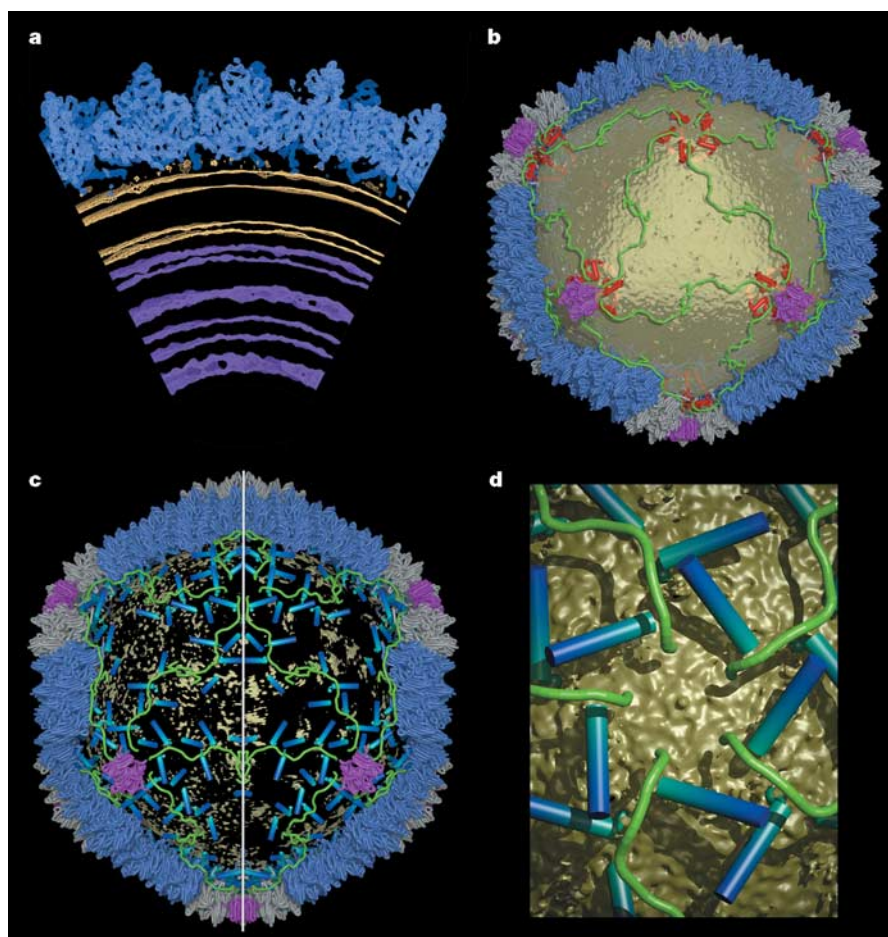


Figure 1 Membrane architecture. **a**, Slice through the cell 2 electron density (calculated as described in Supplementary Fig. 2). The capsid (blue) is contoured at 2σ (relative to the mean), lipid headgroups (yellow) at 0.4σ and DNA layers (magenta) at 0.1σ . **b**, Partial removal of the PRD1 capsid exposes the membrane (yellow). Peripentonal trimers of P3 are coloured grey, all others blue. P31 (refs 3, 8) pentamers at the vertices are shown in magenta (one has been removed to show the continuous curvature of the membrane over the icosahedral five-fold axes). The five copies of P16 (ref. 3) at each vertex are coloured

red, and the P30 network around the facets is shown in green^{3,7}. **c**, Composite of electron density for the membrane outer leaflet for cell 2 (right, at 1.2σ) and cell 1 (left, at 1σ)³. Map calculation is described in Supplementary Fig. 2. Cell 1 is shown as the mirror image of cell 2. P3 N-terminal helices are shown as blue ramped cylinders contacting the membrane at their cyan ends. **d**, Electron density for the membrane outer leaflet at the capsid vertices, showing texture indicative of lateral order.

good model for the crystalline virus), the mean numbers of electrons per molecule for PE and non-PE-type lipids are about 490 and about 650, respectively (see Supplementary Methods). To relate these data quantitatively to the structure, the X-ray data were placed on an absolute scale (see Methods). Figure 2a shows the mean electron density profile for PRD1, calculated so as to reflect the icosahedral geometry of the membrane precisely (see Methods). The difference between the headgroup regions suggests an asymmetrical distribution of headgroup species between the leaflets. To test this, envelopes were calculated defining the inner and outer leaflets of the membrane (see Methods). The inner leaflet contains about 5,517,000 electrons, corresponding to about 11,300 PE lipids. The outer leaflet contains 9,251,000 electrons, equivalent to about 13,200 non-PE and about 1,900 PE-type molecules, which satisfies the constraint that the number of PE and non-PE type lipids be approximately equal. The asymmetrically distributed electron density for the membrane can therefore be completely accounted for by about 25,900 lipids, which is consistent with estimates based on published data on lipid volumes²⁰ that predict the number to be about 25,300 (see Supplementary Methods). Because the relative proportions of PE and phosphatidylglycerol in the host and viral membranes differ¹⁹, the asymmetric headgroup distribution in the viral membrane is probably established during assembly rather than inherited from the host cytoplasmic membrane. Some redistribution of lipids may in addition accompany the roughly 5% radial expansion of the membrane after DNA packaging^{5,6}. This expansion would be expected to lead to a 10% thinning of the membrane, in line with the 32 Å peak-to-peak distance between the headgroups in the electron density profile, which would be about 36 Å in an unstrained membrane of this fatty acid composition²⁰. This lipid asymmetry is probably functionally significant, because there is experimental evidence that negatively charged objects and zwitterionic lipid headgroups experience an electrostatic attraction²¹. Accordingly, the outermost DNA layer is closely associated with the inner leaflet of the membrane (peak-to-peak distance about 20 Å in the profile). Furthermore, the negative charge of the outer leaflet might assist capsid assembly around the membrane because the underside of the P3 trimer is positively charged (S. D. Benson, J. K. H. Bamford, D. H. Bamford and R. M. Burnett, unpublished data). In addition to this global polarization there is lateral segregation of lipid headgroups within the outer leaflet of the membrane (Fig. 1c). This suggests that the three principal types of lipid found in the outer leaflet (PE, phosphatidylglycerol and cardiolipin) are

not distributed uniformly, and it seems plausible that interactions with proteins may orchestrate their organization.

In many dsDNA bacteriophages, the DNA is wound as a spool around an axis running through the packaging vertex²². Presumably the PRD1 genome is organized in the same way, because the virion possesses a unique packaging vertex^{9,10} and its electron density map displays the same concentric DNA rings as those seen in icosahedrally averaged cryo-electron microscopy (cryo-EM) reconstructions of other dsDNA phages, for example filled T7 heads²². The spacing between the second, third and fourth DNA layers is about 22 Å, corresponding to hexagonal packing with an interaxial spacing of about 26 Å. Although the lateral DNA density in the first layer is the same as in the second, the increased spacing between them (about 26 Å) indicates a helix misalignment (Fig. 2b), reducing the DNA density by a factor of $\cos(30^\circ)$ over that in the bulk (110–165 Å radius in Fig. 2a). This might be caused by specific interactions between the outermost DNA layer, the viral membrane and internal domains of proteins such as P16, which penetrate the membrane. A bulk DNA concentration of 560 mg ml⁻¹ would be required to give this inter-helical spacing²³ between the second, third and fourth DNA layers. However, the average DNA concentration in PRD1, calculated from the volume enclosed by the membrane, is about 430 mg ml⁻¹. Because DNA is relatively inflexible there is probably very little DNA within about a 110 Å radius of the particle centre, raising the effective DNA concentration elsewhere (110–200 Å radius) to about 515 mg ml⁻¹, which would account for the entire PRD1 genome (9.7 MDa). Comparisons with the packaging studies on bacteriophage ϕ 29 (ref. 24; see Methods) yield an estimate for the DNA pressure of about 45 atm for PRD1. This might underlie the apparent compression of the DNA layers and puts the membrane under strain in the packaged virion (beyond that which it would be expected to sustain²⁵). This strain is presumably balanced by interactions with the protein capsid. The maximum membrane tension will be at the vertices where the radius of curvature is the smallest, and presumably this is a source of metastability. The intimate interaction between the membrane and the highly pressurized viral genome suggests how the initial stage in ejection of the DNA by means of a membrane tube might be coordinated.

The detailed nature of the present crystallographic analysis reveals unusual aspects of the organization of the membrane, ranging from local structure of the headgroups to lateral segregation within a leaflet and segregation between the leaflets. It is unlikely

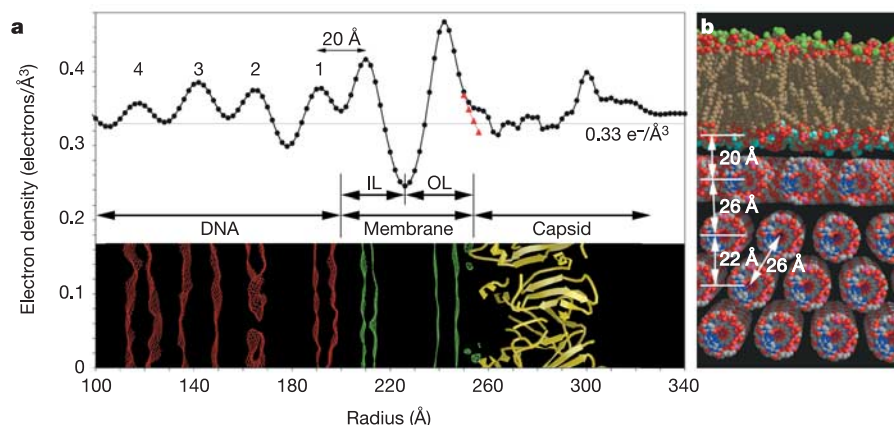


Figure 2 Membrane–DNA interactions. **a**, Electron density profile (black, circles). Distances are measured from the particle centre along the icosahedral three-fold axis. The red line shows the fall-off in the electron density with protein contributions excluded, which allowed the boundary of the outer leaflet to be defined as the surface over which the mean electron density was $0.33 \text{ e}^- \text{Å}^{-3}$. IL and OL mark the boundaries for the inner and

outer leaflets, respectively. DNA layers are numbered 1 to 4. **b**, Space-filling representation of the membrane and DNA based on measurements from the electron density profile, showing the disruption in helix packing between the first and second DNA layers. Glycerol and ethanolamine headgroups are coloured green and cyan, respectively, and phosphate groups are shown in red.

that these aspects are unique to PRD1; for instance, many membrane systems are thought to be asymmetrical^{26,27}. The structure of PRD1 exemplifies how the physicochemical properties of membranes are exploited in unexpected ways in biological systems. □

Methods

Putting the cell 2 map on an absolute scale

Density values in the cell 2 map, ρ , are related to the absolute electron density ρ_0 in units of electrons per $\text{\AA}^3$ ($e^- \text{\AA}^{-3}$) by means of the equation $\rho = k(\rho - \bar{\rho})$, where k is a scale factor and $\bar{\rho}$ is the mean electron density in the unit cell. The absolute value of the electron density in the solvent regions of the crystals was estimated to be $0.344 e^- \text{\AA}^{-3}$ (see Supplementary Methods). The atomic model of P3 (ref. 3) with all hydrogen atoms removed (containing 87.1% of the full complement of 16,327,200 electrons) was solvent corrected to $0.300 e^- \text{\AA}^{-3}$ to give the appropriate contrast with the protein, the edge of the solvent mask was smoothed ($B_{\text{sol}} = 100 \text{\AA}^2$), and a set of structure factors was calculated with program X-PLOR 3.1 (ref. 28). All pixels in the cell 2 map farther than $2.5 \text{\AA}$ from atoms in the P3 model were flattened to the solvent level (program GAP; D.I.S./J.M.G., unpublished program). Wilson scaling of structure factors calculated by inversion of this map, using data between 8.5 and $6 \text{\AA}$, gave $k = 14,845$ (ISO_SCALE; D.I.S./J.M.G., unpublished program), which was scaled down to $12,930$ to account for the absence of hydrogen in the model. The mean electron density associated with pixels in the P3 mask (volume $41,812,500 \text{\AA}^3$) was $0.391 e^- \text{\AA}^{-3}$ and the corresponding mean density in the cell 2 map was 623 . This gave $\bar{\rho} = 0.342 e^- \text{\AA}^{-3}$. The bulk solvent level in the cell 2 map was -81 , which gave a new estimate of the absolute value of the solvent electron density of $0.336 e^- \text{\AA}^{-3}$. The above procedure was repeated with this solvent electron density, yielding $k = 12,864$ and $\bar{\rho} = 0.342 e^- \text{\AA}^{-3}$, which predicted a new solvent electron density of $0.336 e^- \text{\AA}^{-3}$, the same as the previous value. These values of k and $\bar{\rho}$ thus give an absolute electron density scale in which the protein has the correct electron density and the solvent level relative to the protein matches that in the cell 2 map.

Calculation of the electron density profile and membrane envelopes

A net of atoms on a square $10 \text{\AA}$ grid was fitted manually into the electron density for the outer-leaflet headgroups to cover the icosahedral asymmetric unit. This net was then linearly interpolated onto a $2.5 \text{\AA}$ grid. This gave a set of coordinates defining the curvature of the membrane, which intersected the icosahedral three-fold symmetry axes at a distance of $242 \text{\AA}$ from the virus centre. A set of operators were defined that scaled these coordinates relative to the particle centre in $2 \text{\AA}$ steps along the three-fold axes. Application of these operators to the coordinates generated a family of nested surfaces throughout the depth of the virus. An envelope was generated for each surface by using spheres of radius $2 \text{\AA}$ around each coordinate. Calculation of the mean electron density inside the envelope for each surface gave an electron density profile for the virus as a function of distance along the icosahedral three-fold axes in a manner consistent with the geometry of the virus, in contrast with the standard radial profiling programs, which assume spherical symmetry. These surfaces were further used to define envelopes for the inner and outer membrane leaflets using the boundaries defined in Fig. 2a. Pixels within $2.5 \text{\AA}$ of any protein C α atom were excluded. The inner and outer leaflets had volumes of $16,339,000 \text{\AA}^3$ and $26,063,000 \text{\AA}^3$ and had average electron densities of 0.338 and $0.355 e^- \text{\AA}^{-3}$, respectively.

DNA density

The internal volume of the prolate $\phi 29$ head was modelled as a cylinder of radius $182 \text{\AA}$ and depth $313 \text{\AA}$, with spherical caps at either end of radius $212 \text{\AA}$, on the basis of measurements from the cryo-EM reconstruction²⁹. The average $\phi 29$ DNA concentration was calculated to be 460 mg ml^{-1} , given a genome mass of about 12.5 MDa . The average PRD1 DNA concentration of 430 mg ml^{-1} is about 90% of that in $\phi 29$ and thus the internal pressure of the PRD1 DNA should correspond to that in $\phi 29$ at 90% packaging²⁴, giving a value of about 45 atm.

Received 9 March; accepted 21 September 2004; doi:10.1038/nature03053.

- Zhang, W. *et al.* Visualization of membrane protein domains by cryo-electron microscopy of dengue virus. *Nature Struct. Biol.* **10**, 907–912 (2003).
- Mancini, E. J., Clarke, M., Gowen, B. E., Rutten, T. & Fuller, S. D. Cryo-electron microscopy reveals the functional organization of an enveloped virus, Semliki Forest virus. *Mol. Cell* **5**, 255–266 (2000).
- Abrescia, N. G. *et al.* Insights into assembly from structural analysis of bacteriophage PRD1. *Nature* doi:10.1038/nature03056 (this issue).
- Bamford, D. H., Caldentey, J. & Bamford, J. K. Bacteriophage PRD1: a broad host range dsDNA tectivirus with an internal membrane. *Adv. Virus Res.* **45**, 281–319 (1995).
- Butcher, S. J., Bamford, D. H. & Fuller, S. D. DNA packaging orders the membrane of bacteriophage PRD1. *EMBO J.* **14**, 6078–6086 (1995).

- San Martín, C. *et al.* Combined EM/X-ray imaging yields a quasi-atomic model of the adenovirus-related bacteriophage PRD1 and shows key capsid and membrane interactions. *Structure* **9**, 917–930 (2001).
- Rydman, P. S., Bamford, J. K. & Bamford, D. H. A minor capsid protein P30 is essential for bacteriophage PRD1 capsid assembly. *J. Mol. Biol.* **313**, 785–795 (2001).
- Rydman, P. S. *et al.* Bacteriophage PRD1 contains a labile receptor-binding structure at each vertex. *J. Mol. Biol.* **291**, 575–587 (1999).
- Strömsten, N. J., Bamford, D. H. & Bamford, J. K. The unique vertex of bacterial virus PRD1 is connected to the viral internal membrane. *J. Virol.* **77**, 6314–6321 (2003).
- Gowen, B., Bamford, J. K., Bamford, D. H. & Fuller, S. D. The tailless icosahedral membrane virus PRD1 localizes the proteins involved in genome packaging and injection at a unique vertex. *J. Virol.* **77**, 7863–7871 (2003).
- Bamford, D. & Mindich, L. Structure of the lipid-containing bacteriophage PRD1: disruption of wild-type and nonsense mutant phage particles with guanidine hydrochloride. *J. Virol.* **44**, 1031–1038 (1982).
- Rydman, P. S. & Bamford, D. H. Bacteriophage PRD1 DNA entry uses a viral membrane-associated transglycosylase activity. *Mol. Microbiol.* **37**, 356–363 (2000).
- Rydman, P. S. & Bamford, D. H. The lytic enzyme of bacteriophage PRD1 is associated with the viral membrane. *J. Bacteriol.* **184**, 104–110 (2002).
- Bamford, J. K. *et al.* Diffraction quality crystals of PRD1, a 66-MDa dsDNA virus with an internal membrane. *J. Struct. Biol.* **139**, 103–112 (2002).
- Grahn, A. M., Daugelavicius, R. & Bamford, D. H. Sequential model of phage PRD1 DNA delivery: active involvement of the viral membrane. *Mol. Microbiol.* **46**, 1199–1209 (2002).
- Davis, T. N., Muller, E. D. & Cronan, J. E. Jr The virion of the lipid-containing bacteriophage PR4. *Virology* **120**, 287–306 (1982).
- Grahn, A. M., Caldentey, J., Bamford, J. K. & Bamford, D. H. Stable packaging of phage PRD1 DNA requires adsorption protein P2, which binds to the IncP plasmid-encoded conjugative transfer complex. *J. Bacteriol.* **181**, 6689–6696 (1999).
- Tuma, R., Bamford, J. K., Bamford, D. H. & Thomas, G. J. Jr Structure, interactions and dynamics of PRD1 virus. II. Organization of the viral membrane and DNA. *J. Mol. Biol.* **257**, 102–115 (1996).
- Laurinavicius, S., Käkälä, R., Somerharju, P. & Bamford, D. H. Phospholipid molecular species profiles of tectiviruses infecting Gram-negative and Gram-positive hosts. *Virology* **322**, 328–336 (2004).
- Nagle, J. F. & Tristram-Nagle, S. Structure of lipid bilayers. *Biochim. Biophys. Acta* **1469**, 159–195 (2000).
- Struth, B. *et al.* Organization of two-dimensional phospholipid monolayers on a gel-forming substrate. *Phys. Rev. Lett.* **88**, 025502 (2002).
- Cerritelli, M. E. *et al.* Encapsidated conformation of bacteriophage T7 DNA. *Cell* **91**, 271–280 (1997).
- Livolant, F. & Leforestier, A. Condensed phases of DNA: Structures and phase transitions. *Prog. Polym. Sci.* **21**, 1115–1164 (1996).
- Smith, D. E. *et al.* The bacteriophage $\phi 29$ portal motor can package DNA against a large internal force. *Nature* **413**, 748–752 (2001).
- Evans, E., Heinrich, V., Ludwig, F. & Rawicz, W. Dynamic tension spectroscopy and strength of biomembranes. *Biophys. J.* **85**, 2342–2350 (2003).
- Beyers, E. M., Comfurius, P., Dekkers, D. W. C. & Zwaal, R. F. A. Lipid translocation across the plasma membrane of mammalian cells. *Biochim. Biophys. Acta* **1439**, 317–330 (1999).
- Casper, D. L. & Kirschner, D. A. Myelin membrane structure at $10 \text{\AA}$ resolution. *Nature New Biol.* **231**, 46–52 (1971).
- Brunker, A. T. *X-PLOR, Version 3.1: A System for X-ray Crystallography and NMR* (Yale University Press, New Haven, Connecticut, 1992).
- Tao, Y. *et al.* Assembly of a tailed bacterial virus and its genome release studied in three dimensions. *Cell* **95**, 431–437 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Sansom and P. Bond (Laboratory of Molecular Biophysics, University of Oxford) for providing coordinates of lipid bilayers, the staff at the European Synchrotron Radiation Facility for beamline support, and E. Mancini for assistance with data collection. This investigation was supported by a research grant from the Academy of Finland to J.K.H.B., research grants and the Finnish Centre of Excellence Program (2000–2005) from the Academy of Finland to D.H.B., the Biotechnology and Biological Sciences Research Council and the Medical Research Council, UK, and a grant from the Human Frontiers Science Program to D.H.B. and D.I.S. J.M.G. is supported by the Royal Society, and D.I.S. by the UK Medical Research Council.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.I.S. (dave@strubi.ox.ac.uk) or D.H.B. (dennis.bamford@helsinki.fi).

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayouni (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Ostrup (4994)

France/ Switzerland:

Amélie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

**Germany/ Austria/ Italy/
The Netherlands/ Belgium:**

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: r.asami@naturejpn.com

naturejobs

Home and away

After the First World War writers and artists moved to Paris in search of a cheap and exciting place to work. They soon found each other — in Gertrude Stein's house, the Shakespeare and Company bookstore and countless cafés. Today's globally mobile scientific community similarly forms homes-from-home abroad.

Expatriate organizations for scientists exist at every scale and level of formality. At one end of the spectrum are groups such as BioBrits. Based in San Diego, this group of about 100 Britons involved in biotechnology do not even have a website. Members simply meet up each month for a beer and a gossip. At the other extreme is the Society of Chinese Bioscientists in America (SCBA), a group so huge it has its own building, staff and a sophisticated web presence (www.scba-society.org).

Whatever their size, expat organizations are important for job seekers. They pool resources, provide networking opportunities and often, as with the SCBA, collate lists of jobs from all over the world. Recruitment organizations are now trying to capitalize on the strengths of these expatriate communities. One of the most extensive of these is Talent Scotland (www.talentscotland.com), which aims to recruit expatriate Scots working in information technology and the life sciences and bring them back home. It seems to be working: so far, more than 28,000 people have registered at the website, and 750 Scots have returned.

More Scottish expats may soon find the lure of home irresistible, as the country has seen a surge in infrastructure investment in the sciences that should continue for the next several years (see page 128). In Paris, the itinerant writers were dubbed 'the lost generation'. To global recruiters, the present generation of expat scientists can now be considered to be well and truly found.

Paul Smaglik
Naturejobs editor



Contents

REGIONS

Building on
Scotland's success p128

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

High road to Scotland

The creation of Dolly the sheep by nuclear transfer at the Roslin Institute focused the world's attention on Scotland, and private funding and investment in public infrastructure may sustain this interest. This is especially true at the three biggest biomedical hubs: Edinburgh, Glasgow and Dundee, which are adding substantial life-science facilities over the next few years.

Ken Snowden, director of biotechnology at Scottish Enterprise — the country's main economic development agency, funded by the Scottish Executive — says that Scotland has seen a 20% growth in life-science jobs over the past five years, with a total of 26,000 employees across all sectors. The country has 550 life-science organizations, including 50 universities.

About half of Scotland's life-science companies are service rather than product oriented. More than 4,000 people are employed at the two biggest contract-research organizations operating in Scotland — Inveresk and Quintiles. Although service companies provide steady jobs, higher-risk drug-development companies, of which Scotland has few, drive growth. "It's good to have a balance," Snowden says.

In Edinburgh that balance is clear in the Little France area, which has seen a massive build-up and move into both translational medicine and technology transfer, with an 800-bed hospital and research facilities on the same site. The medical school and The Royal Infirmary have already moved there, and in May next year, 600 scientists will begin to fill the new Research Institute for Medical Cell Biology. This will effectively double the number of researchers on site and bring together five key centres of interdisciplinary research in inflammation, infectious diseases, reproductive medicine, neuroscience and post-genomic medicine.

The University of Edinburgh's Little France campus (below) will be the largest biomedical research complex in Europe, says Peter Ghazal.



Picture-postcard scenery belies Scotland's reputation as a centre of scientific and medical excellence.

Little France is stealthily aimed at overtaking regions in England, such as Cambridge, in terms of sheer size and number of researchers — with more than twice as much room as Cambridge's science park. "The Little France cluster will be one of the largest biomedical research complexes in Europe," says Peter Ghazal, director of the Scottish Centre for Genomic Technology and Informatics, housed at the site.

Ghazal says that the site's real strength will be its ability to commercialize technology born of basic research and tested in the clinic. He has experience of that ethos after spending much of his career at the Scripps Research Institute in La Jolla, California. Ghazal is already involved in two spin-out companies: Arrayjet, which makes a microarrayer, and Lab901, which is miniaturizing routine lab tests. He thinks that teaming life-science and physical-science researchers with engineers, then asking physicians what kind of technologies they need, will help to produce more companies.

Not to be outdone by Edinburgh, Glasgow is in the middle of a five-year, £185-million (US\$338 million) building boom. The Glasgow Biomedical Research Centre will open next year, as will a £12-million cardiovascular centre. And a £10-million new building for the Beatson Institute for Cancer Research will open in 2006.

Peter Holmes, vice-principal for research at Glasgow University, says these buildings are creating jobs. He anticipates about 30 new professorial appointments in science over the next year.

Dundee will also see major infrastructure additions and new jobs. The university will open a clinical research centre next year, with an additional 50–60 research jobs. The centre will focus on running academic clinical trials for much of Scotland, and, indeed, the rest of Britain, says Jill Belch, head of vascular medicine. The focus will be on imaging, with a positron emission tomography



C. PHILIP/CORBIS



(PET) scanner, cyclotron and molecular-imaging equipment, and the aim will be to develop non-invasive diagnostics using the technology.

Dundee University's Centre for Interdisciplinary Research, set to open in July 2005, is aimed at drug discovery and development. It will have capacity for compound screening and will emphasize 'orphan tropical diseases', such as African trypanosomiasis, where there is less interest from big drug companies, says Mike Ferguson, professor of molecular parasitology. The £17.5-million building will eventually house about 250 researchers, with 100

scientists, graduate students and postdocs to be recruited. Expanding areas will include medicinal chemistry, computational chemistry and high-throughput screening.

The question remains whether public investment can create private-sector jobs. For service-based companies, the answer is yes. The government-created Scottish Biomedical, in Glasgow's science park, has been so successful that its management bought it back. The contract-research organization initially paid university-based researchers to find and validate drug targets, while the customer retained intellectual-property rights. But demand for its services, particularly from Asian companies looking to get into the British or European markets, has grown so much that the company now has more than 40 in-house scientists, with plans to add more as the company's contracts with university-based scientists expire. And a new, smaller Dundee company, CXR Biosciences, is looking to create a similar success story by making models to detect drug toxicity.

Transforming Scotland into a centre for drug development rather than services is a taller order. But this evolution may be underway. Halfway between Edinburgh and Glasgow, Organon Research Scotland provides an early example. Since 1998 it has shifted from primarily a production facility to a major R&D site — more than tripling the number of scientists. And Ardana Bioscience, an Edinburgh company formed in 2001 to commercialize research from the UK Medical Research Council's Human Reproductive Sciences Unit, already has one product on the market. The company Cyclacel, a University of Dundee spin-out, emphasizes small-molecule drugs and hopes to use the country's strengths in biomarkers to find early signs of drug effectiveness or toxicity.

The growth of these companies, coupled with new academic infrastructure and positions, may give Scotland the balance it needs to capitalize on its long-standing tradition of excellence in medicine.

Paul Smaglik is editor of *Naturejobs*.

Scottish stem cells

The Roslin Institute near Edinburgh and the University of Edinburgh's Institute for Stem Cell Research are set to be the seeds of a Scottish stem-cell empire.

The Roslin supports four embryonic stem-cell programmes, with funding from the government, the US biotech company Geron and Dundee-based CXR Biosciences. The groups are creating new lines, exploring the use of stem cells for toxicity testing and bone and cartilage repair and studying how stem cells grow and differentiate into other cell types. "We've probably got one of the biggest groups in the United Kingdom working on embryonic stem cells," says Harry Griffin, Roslin's director.

But there's much more on the drawing board, says Griffin. Roslin, together with the University of Edinburgh's veterinary school, the Moredun Research Institute and the Scottish Agricultural College, is hoping to create an Edinburgh 'centre of excellence' with some 500 staff. Even more urgent is the need for a stem-cell production facility that meets the quality and safety standards of the European Union (EU), as Griffin and Scottish colleges are keen to create stem

cells that could be used to treat humans. The Scottish Stem Cell Network, a coalition of public and private organizations, is helping to push both projects forward.

The Institute for Stem Cell Research is a major node in the Stem Cell Network. The institute's own research activity is funded at £71 million (US\$130 million) over the next three years, says the director, Austin Smith, and he anticipates that the 80 or so researchers working there now will be 120 in a few years. The institute is also the headquarters for EuroStemCell, a European research collaboration which has received an EU grant of €11.9 million (US\$15 million) over four years for 27 principal investigators, six of whom are in Edinburgh. The building also holds the company Stem Cell Sciences, which aims to provide stem-cell culture and distribution on an international scale.

P.S.

Scottish Stem Cell Network

♦ www.sscn.co.uk

EuroStemCell

♦ www.eurostemcell.org